# Multiscale modeling of DNA, from double-helix to chromatin

Sam Meyer

▶ **To cite this version:**

Sam Meyer. Multiscale modeling of DNA, from double-helix to chromatin. Other [cond-mat.other]. Ecole normale supérieure de lyon - ENS LYON, 2012. English. NNT : 2012ENSL0744 . tel-00756315

**HAL Id: tel-00756315**

**https://theses.hal.science/tel-00756315**

Submitted on 22 Nov 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# THÈSE

en vue de l'obtention du grade de

**Docteur de l'École Normale Supérieure de Lyon - Université de Lyon**

**Discipline : Physique**

**Laboratoire de physique de l'ENS Lyon – Centre Blaise Pascal**
**École Doctorale de Physique et d'Astrophysique (PHAST)**

présentée et soutenue publiquement le 28 septembre 2012

par M. Sam MEYER

---

# Multiscale modeling of DNA, from double-helix to chromatin

---

*Directeur de thèse :* *M. Ralf EVERAERS*
*Co-directeur de thèse :* *M. Richard LAVERY*

*Après l'avis de :* *M. Helmut SCHIESSEL*
*M. Jean-Marc VICTOR*

*Devant la commission d'examen formée de :*

*M. Ralf EVERAERS, ENS Lyon, directeur*
*M. Richard LAVERY, Université Claude Bernard Lyon I, co-directeur*
*M. Michel PEYRARD, ENS Lyon, président*
*M. Helmut SCHIESSEL, Universiteit Leiden, rapporteur*
*M. Jean-Marc VICTOR, Université Pierre et Marie Curie, rapporteur*
*M. Stefan DIMITROV, Université Joseph Fourier Grenoble I, membre*

# Résumé

## Modélisation multi-échelle de l'ADN, de la double-hélice à la chromatine

Dans le noyau des cellules eucaryotes, l'ADN s'enroule autour d'histones pour former des nucléosomes, lesquels s'arrangent à leur tour en une fibre compacte et dynamique appelée chromatine. Les propriétés physiques de cette fibre aux différentes échelles, depuis la double-hélice de l'ADN jusqu'aux chromosomes micrométriques, sont essentielles aux mécanismes complexes de l'expression des gènes et sa régulation. La présente thèse est une contribution au développement de modèles physiques capables de relier les différentes échelles, et d'interpréter et d'intégrer des données provenant d'une large gamme d'approches expérimentales et numériques.

En premier lieu, nous utilisons des simulations de dynamique moléculaire d'oligomères d'ADN pour étudier l'ADN double-hélical à différentes températures. Nous estimons la contribution séquence-dépendante de l'entropie à l'élasticité de l'ADN, en lien avec des expériences récentes sur la longueur de persistence de l'ADN.

En second lieu, nous modélisons les interactions ADN-histones au sein du Nucleosome Core Particle. Nous utilisons la nanomécanique de l'ADN afin d'extraire un champ de force d'un ensemble de structures cristallographiques du nucléosome et de données de dynamique moléculaire.

En troisième lieu, nous étudions la partie plus molle du nucléosome, l'ADN linker entre les core particles, qui s'associe transitoirement à l'histone H1 pour former un "stem". Nous combinons des informations structurales existantes avec des données expérimentales à deux résolutions différentes (DNA footprinting et électro-microscopie) afin de développer un modèle de stem à l'échelle nanométrique.

**Mots-clés :** ADN, chromatine, simulation numérique, coarse-graining

# Abstract

## Multiscale modeling of DNA, from double-helix to chromatin

In the nucleus of eukaryotic cells, DNA wraps around histone proteins to form nucleosomes, which in turn associate in a compact and dynamic fiber called chromatin. The physical properties of this fiber at different lengthscales, from the DNA double-helix to micrometer-sized chromosomes, are essential to the complex mechanisms of gene expression and its regulation. The present thesis is a contribution to the development of physical models, which are able to link different scales and to interpret and integrate data from a wide range of experimental and computational approaches.

In the first part, we use Molecular Dynamics simulations of DNA oligomers to study double-helical DNA at different temperatures. We estimate the sequence-dependent contribution of entropy to DNA elasticity, in relation with recent experiments on DNA persistence length. In the second part, we model the DNA-histone interactions within the nucleosome core particle, using DNA nanomechanics to extract a force field from a set of crystallographic nucleosome structures and Molecular Dynamics snapshots.

In the third part, we consider the softer part of the nucleosome, the linker DNA between core particles which transiently associates with the histone H1 to form a "stem". We combine existing structural knowledge with experimental data at two different resolutions (DNA footprints and electro-micrographs) to develop a nanoscale model of the stem.

**Keywords :** DNA, chromatin, numerical simulation, coarse-graining

# Résumé détaillé

## Modélisation multi-échelle de l'ADN, de la double-hélice à la chromatine

Dans le noyau des cellules eucaryotes, l'ADN s'associe avec des protéines pour former une fibre compacte et dynamique, la chromatine. La structure élémentaire de cette fibre, appelée nucléosome, est formée d'environ 150 paires de bases (pb) d'ADN, enroulées autour d'un octamère d'histones. Les propriétés physique de la fibre de chromatine sont essentielles aux mécanismes complexes d'expression des gènes et de sa régulation, et elles sont l'objet d'un effort de recherche intense depuis plusieurs décennies.

Des méthodes expérimentales variées ont été mises en oeuvre : études structurales de haute résolution (cristallographie aux rayons X, RMN), méthodes d'imagerie (électromicroscopie, AFM...), mesures mécaniques sur particules uniques (pinces optique et magnétique), mesures structurales indirectes (méthodes biochimiques, FRET...). Ces différentes méthodes fournissent ainsi des informations de différentes natures, et à différentes résolutions spatiales qui ne se recouvrent pas toujours. Le développement de modèles physiques permet de relier ces différentes échelles, et d'interpréter et d'intégrer les données expérimentales de sources différentes. La présente thèse est une contribution au développement de tels modèles.

Dans un premier temps (chapitre 0), nous présentons des généralités sur (i) le système biologique étudié, aux différentes échelles ; (ii) les modèles physiques et (iii) les méthodes numériques utilisés dans les chapitres suivants. Le travail de thèse lui-même est divisé en trois chapitres, correspondant à trois échelles spatiales successives dans l'organisation de la chromatine.

Dans le chapitre 1, nous étudions la dépendance en température de la flexibilité de la double-hélice d'ADN. Nous effectuons des simulations de dynamique moléculaire de 4 18-mères d'ADN de différentes séquences, pour des températures entre 273K et 350K. L'analyse des trajectoires fournit des informations détaillées sur les fluctuations intra- et inter-pb. Dans le premier cas, la contribution entropique est considérable, en particulier dans les directions stretch (séparation des bases) et opening (rotation dans le plan de la pb) : la dénaturation de la double-hélice est ainsi probablement initiée par des fluctuations importantes selon ces deux directions, jusqu'à la rupture coopérative des pb et la formation de bulles de dénaturation (phase de prémelting). La contribution entropique dans la flexibilité inter-pb est plus faible, mais néanmoins détectable. En utilisant des méthodes de coarse-graining, nous montrons qu'elle fournit une contribution linéaire en température à la rigidité à large-échelle de l'ADN (longueur de persistence), d'amplitude comparable à celle due aux bulles de d enaturation dans une gamme de températures 10°-50°C.

Dans le chapitre 2, nous étudions les interactions entre les histones et l'ADN enroulé dans le nucléosome. Nous montrons que la connaissance (i) de la mécanique de l'ADN et (ii) d'un

nombre suffisant de structures haute-résolution du nucléosome permet d'en extraire un potentiel d'interaction histone-ADN, dans l'approximation élastique linéaire. En effectuant cette extraction sur un ensemble de données cristallographiques et de snapshots de dynamique moléculaire, nous trouvons le résultat surprenant que (i) aux points d'ancrage, les forces exercées par les histones sont souvent répulsives ; (ii) les interactions estimées correspondent à des potentiels harmoniques instables. Pour interpréter ces premiers résultats, nous faisons l'hypothèse (pour le moment spéculative) que l'ADN nucléosomal serait dans un état de tension métastable, à la manière d'un élastique tendu autour d'un objet cylindrique, où ce sont ses propriétés élastiques qui le maintiennent en place, plutôt que les (ou en plus des) interactions attractives de l'octamère.

Le chapitre 3 est consacré à la détermination de la structure du "stem" nucléosomal, c'est-à-dire la jonction des deux branches d'ADN externes du nucléosome (linker) en présence du linker histone H1. Cette structure joue un rôle déterminant dans la structuration de la fibre de chromatine, mais elle n'avait pu être déterminée en raison du caractère fluctuant de cette partie du nucléosome. Nous combinons des données existantes sur la structure du nucléosome et l'élasticité de l'ADN avec deux types de données expérimentales obtenues par des groupes de biologistes (D. Angelov, S. Dimitrov, J. Bednar). L'analyse de gels de protection de l'ADN contre des attaques de radicaux hydroxyles fournit des informations sur les contacts moléculaires, à une résolution de 1pb. Elle permet de tester différents modèles proposés pour le placement de H1, et d'en tirer une structure de stem de protection maximale. Les images de cryo-microscopie électronique ont une résolution plus faible ; la combinaison des deux techniques permet de construire un modèle de stem incluant les fluctuations thermiques. Nous estimons que la présence de H1 réduit considérablement la liberté conformationnelle du nucléosome, en imposant une structure rigide de $20\pm2$ pb sur chaque linker. Nous discutons ensuite les conséquences de notre modèle pour la fibre de chromatine. L'analyse d'un gel de dinucléosomes montre que l'interaction entre stems successifs peut engendrer des déformations de leur structure, mécanisme qui pourrait également exister dans la fibre in vivo.

# Table des matières

# Introduction

DNA is the common substrate of the genomic information in all living organisms. Its molecular structure is known since 1953 [Watson and Crick 1953] : a long double-helix, where the *genetic sequence* is encoded in a succession of nucleotides, which act as a four-letter alphabet. In these last years, technical progress allowed the deciphering of entire genomes, including that of the human species [Collins et al. 2004]. Has our understanding of genetics and of its physical basis reached its ultimate stage ?

As a matter of fact, in spite of these discoveries, we understand little. The knowledge of the sole molecular structure of DNA and its sequence does not account for the mechanisms of genetic expression, its regulation, or DNA replication, at the molecular level, and even less so at larger scales. A stunning illustration of this problem is the cellular differentiation : two cells of identical genome, placed in different conditions, express different genes and evolve differently in an irreversible way. The interaction between the DNA and the cellular environment is not one-way, but rather a complex set of relations. The environment does not modify the genome itself, but it may influence the *packaging* of its molecular carrier at different lengthscales, hence the variability of genetic expression [Calladine and Drew 1997]. This organization is the subject of the present study.

The human genome is made of approximately 3 billion base-pairs (bp), divided in 46 pieces of DNA, for a total molecular length of around 2 meters. In physiological conditions, the coil formed by each of these pieces has a diameter of around 100 microns [Schiessel 2003], and yet it must fit into the micrometer-sized nucleus of each of our cells. DNA is therefore strongly confined *in vivo*.

To be transcribed or replicated, the molecule must however be accessible to proteins, and consequently, the spatial crowding must be limited in the expressed region. Thus, the cell needs a dedicated *machinery*, which achieves the contradictory features of ensuring an important global compaction of the molecule, while allowing for local openings when required. This property is quite analogous to that of a more familiar system : a library [Schiessel 2003]. Here, an important amount of information is stored in closed books disposed on shelves. It is thereby stored in a limited space, as compared to the total length of the text contained in the library. But accessing some information requires the opening ("decompacting") of the appropriate book.

Depending on its state, and that of the genome organizing machinery, each cell expresses some particular subset of "open" genes : this mechanism plays a major role in cellular differentiation. However, the analogy of the library raises some difficulties when considering not only the compaction, but also the accessibility of this information and its *regulation*. The librarian can influence the reader's choice by disposing a given set of books openly on display, and by storing away in some secluded corner the ones considered as inadvisable. If the former ones were previously stored away, he has an index to find them immediately when desired. In

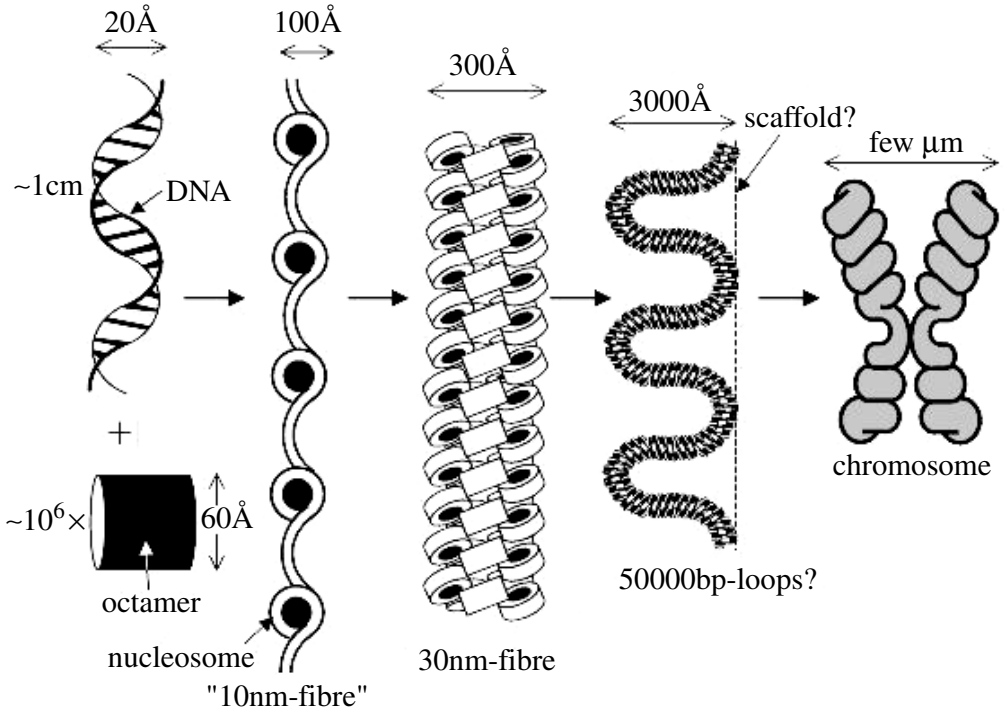


FIGURE 1 – Schematic depiction of the successive organization levels of the chromatin structure

the cell, how does the machinery "know" which region to unfold, depending on the particular needs at a given time ? The answer to this question remains largely unknown.

In eukaryotic cells, the machinery is called *chromatin*. It is a complex hierarchical structure resulting from the association of an important amount of proteins with DNA, as schematically depicted on Fig. 1. Its primary structure is made of $\sim 200$ base-pair (bp) of DNA wrapped around a group of proteins, and called the *nucleosome*. In a solution of low ionic strength, the chromatin fiber takes the form of a sequence of these nucleosomes, with an diameter of $\sim 10$ nanometer (nm). *In vivo*, it folds into a secondary structure, of around 30 nm diameter. At some stages of the cellular cycle, the latter may form a tertiary, or even quaternary structure, strongly condensing into dense *mitotic chromosomes*, visible in the optical microscope [Woodcock and Dimitrov 2001].

The accessibility of the DNA sequence to the proteins which express the genes is largely conditioned by the state of the chromatin fiber [Wu et al. 2007]. Changes in its structure, probably related to chemical modifications with physical consequences on the nucleosomes, play a major role in different aspects of genetic regulation, and subsequently on the cellular cycle, but probably also for instance on aging [Woodcock and Dimitrov 2001] and cancer [Michor et al. 2011]. These modifications are induced *in vivo* by a whole group of specialized *chromatin remodeling* proteins, which are essential to the proper packaging of the DNA throughout the cell cycle [Flaus and Owen-Hughes 2001].

Due to its biological importance, chromatin has been the object of intense research for decades, and in particular since the discovery of the nucleosome in 1973 [Kornberg 1974]. In the last 15 years, nucleosome crystals revealed their internal structure at almost atomic resolution. New experimental methods have been developed, which increase the existing and potential amount of available information. A non-exhaustive list of these methods includes, approxima-

tely ordered by spatial scale :

- **<1nm** : high-resolution structural techniques : X-ray crystallography, Nuclear Magnetic Resonance (NMR), neutron scattering
- **1-10nm** : bulk and single-molecule Fluorescence Resonance Energy Transfer (FRET), biochemical mapping
- **10-100nm** : single-molecule manipulation experiments (with optical or magnetic tweezers), high-resolution imaging : Cryo-Electron Micrograph (CEM), Atomic Force Microscopy (AFM)
- **>100nm** : large-scale distance mapping methods, for instance Fluorescence In-Situ Hybridization, optical microscopy

In addition of these experimental sources of knowledge, one should mention the rapid development of numerical simulation-based data, in particular all-atomic Molecular Dynamics (MD) for molecular structures. This technique needs to be carefully calibrated and validated by comparison to experiments, but its increasing reliability and accuracy make it an extremely valuable source of information.

And yet, in spite of this remarkable progress, the structural and physical properties of the chromatin fiber at different lengthscales have remained largely elusive. Above the nucleosome scale, the fiber is intrinsically soft, which opposes the usual high-resolution techniques. Its density is a major obstacle for imaging techniques, which do not reach the sufficient resolution to capture the details of nucleosomal folding. Structural information thus remains sparse. Another source of difficulties is that the fiber presents a natural variability, and the preparation of well-defined samples is delicate. The results obtained from indirect structural studies are therefore sometimes contradictory [Schalch et al. 2005, Dorigo et al. 2004, van Holde and Zlatanova 2007, Wu et al. 2007, Robinson et al. 2006].

An important obstacle is that all the mentioned sources of information are different in lengthscale, but also in the kind of information they provide, which can therefore not always be directly compared and integrated. To overcome this obstacle requires the introduction of *physical models*. The latter can bridge data of different resolutions by computing coarse-graining relations. They allow a quantitative structural interpretation from indirect data : the most straightforward example is the inference of fiber structural models from force-extension curves. Most importantly, they provide not only structural information, but also the physical mechanisms governing complex physiological processes.

In this work, we develop three examples of such models, at three successive lengthscales. They are grouped in three chapters :

1. We run and analyze MD trajectories of DNA oligomers at different temperatures, and estimate the contribution of entropy to the double-helical elasticity, at the base-pair and base-pair step level. We then use coarse-graining relations to relate the computed values to the persistence length of the molecule, which has been recently measured over a wide range of temperatures [Geggier et al. 2011]. Entropy yields two contributions to the apparent persistence length : by decreasing the stiffness of the double-helix, where the contribution grows linearly with temperature, and by local openings in the premelting regime [Theodorakopoulos and Peyrard 2012], where it is strongly nonlinear. We show that in the temperature range 10°C < T <50 °C, the two contributions are of the same order, and together, are compatible with the experimental datapoints.
2. We develop a method to extract the potential for the histone-DNA interactions in the

nucleosome from a set of crystallographic structures and MD snapshots, and discuss the first results.

3. We combine existing high-resolution structural models with the results of footprinting experiments and cryo-electron-micrographs to build a nanoscale model of the stem structure formed by the nucleosome in presence of the linker histone H1 [Bednar et al. 1998]. We show that the binding of H1 considerably reduces the conformational freedom of the nucleosome, and results in an approximately rigid structure of $187\pm4$ base-pair, with an experimentally observable effect extending at least 20 base-pair beyond on either side. We then discuss the consequences of our stem modeling for the chromatin fiber. The analysis of a dinucleosome footprinting gel shows that the stem interaction results in a deformed structure, a mechanism that could also exist in the fiber *in vivo*.

# Chapitre 0

# Preliminaries

This introductory chapter aims to lay down a body of knowledge, which will be referred to throughout this work. Three sections are devoted to :

1. **biological system** : we describe structural and some biological aspects of DNA and chromatin
2. **physical models** : we present the physical models that will be used in this work : the atomistic force fields, and two elastic models of DNA corresponding to two different lengthscales : the rigid base-pair level and the worm-like chain (wlc) model, as well as coarse-graining relations between these successive levels of description
3. **simulation methods** : we introduce the numerical methods that we will be using for simulating the considered models : Molecular Dynamics (MD) and Monte Carlo (MC) sampling.

The different sections are independent of one another, and so are the upcoming chapters : the reader may read only a single section and "pick" the relevant information for a given chapter. Here, we present only general aspects of the different topics ; more specific ones are described directly in the different chapters.

## 0.1 Biological system : from DNA to chromatin

### 0.1.1 Molecular structure of DNA

Deoxyribonucleic acid (DNA) is a heteropolymer of 4 elementary building blocks called *nucleotides*, which form two separate strands [Calladine and Drew 1997]. In physiological conditions, the strands associate by forming hydrogen bonds and inter-base stacking interactions, resulting in the well-known "B-DNA" double-helix (dh) structure first described in 1953 [Watson and Crick 1953] : see Fig. 0.1 A and B.

The *nucleotide* is a compound of three elements : a phosphate group $PO_4^-$, a sugar ring (a five-carbon cyclic group) and a (nucleo)base, which is a complex organic group containing one or two cycles. The *backbone* of each strand is made of the succession of covalently bound phosphate groups and sugars (Fig. 0.2A). The contacts of the sugar groups are asymmetric : we therefore set a conventional direction of reading, quoted 5'$\rightarrow$3', which refers to the terminal carbon atom on either side. Note that the orientations of the two strands are antiparallel : the end of one strand is spatially close to the beginning of the other strand.

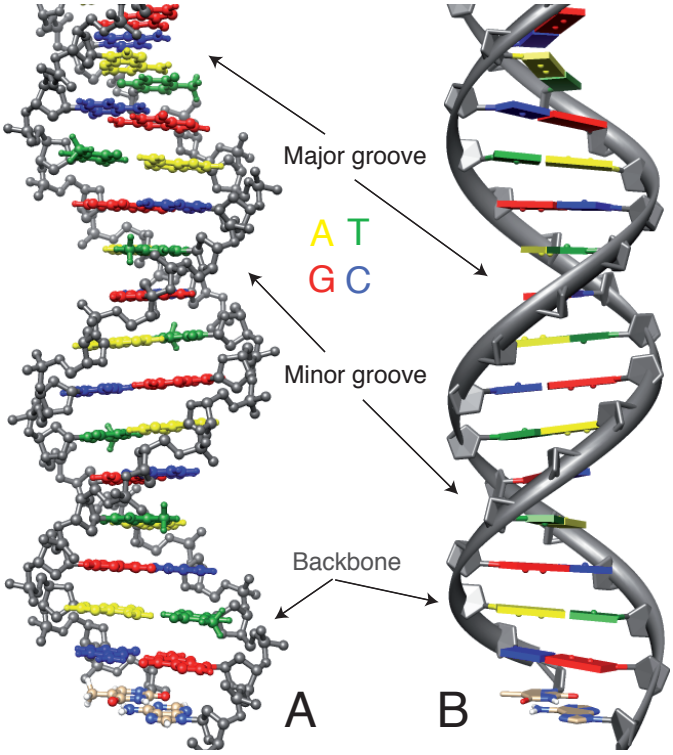


FIGURE 0.1 – dsDNA in the B-DNA conformation : all-atom (A) and schematic (B) representation. The backbones are shown in gray, and the bases are colored. Pictures generated with Chimera Pettersen et al. [2004]

In aequous solution, the phosphate groups dissociate, which makes DNA a strongly charged polyelectrolyte, with 2 negative charges per base-pair (bp). This results in important electrostatic effects, which contribute to the physical properties of the molecule in a complex and yet unresolved way. In particular, the solvation of the molecule in physiological solution where the salt concentration is important ($\sim 0.15$ mol/L), involves a structured aggregation of counterions around the molecule, which partly neutralize the phosphates, and thus probably modify the mechanical properties by reducing the electrostatic repulsion between them. Let us mention that in such conditions, the electrostatic interactions are screened and decay rapidly, with a Debye screening length of $\sim 1$ nm [Gelbart et al. 2000]. The nucleotides differ by the bases, which can be of four types belonging to two groups :

- *purines* A (adenine) and G (guanine) have two organic cycles
- *pyrimidines* T (thymine) and C (cytosine) have only one

In dsDNA, these hydrophobic bases are buried inside the molecule, and associate specifically by either two (A-T) or three (C-G) hydrogen bonds, forming a rather flat complex called bp : see Fig. 0.2B. The two purine-pyrimidine complexes have similar sizes, hence the approximately regular dh structure. The *genetic sequence* of a given strand is simply the sequence of the bases, read in the conventional direction. A consequence of the base-pairing specificity is that the sequences of the two complementary strands are univocally related, which "immediately suggests a possible copying mechanism for the genetic material" [Watson and Crick 1953].

In the dh, the backbones are separated by two helically-shaped grooves of different widths, the major groove and the minor groove. Depending on the physical and chemical conditions, the dh can adopt other conformations than B-DNA, where in particular the groove sizes are different. In this work, we will restrict to the physiologically relevant B-DNA conformation and fluctuations around it. Let us only mention the name of an alternate conformation : A-DNA, which is relevant in solution of low ionic force.

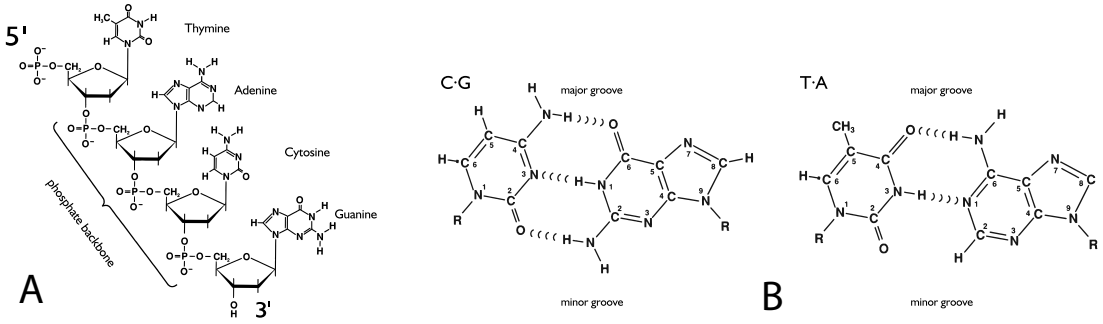


FIGURE 0.2 – Chemical composition of dsDNA : (A) Watson-Crick base-pairing ; (B) Single strand. Adapted from [Sinden 1994]

The dh has a diameter of around 2 nm, and the helical pitch is of 10.5 bp, *i.e.* two successive base-pairs are rotated by a twist angle of around 34°. Note that the latter value is an average, and exhibits some static variation depending on the sequence, as well as thermal fluctuations which distort the dh from the ideal model. These distortions remain rather limited at room temperature however, the molecule being rather stiff, with a persistence length of $\sim$50 nm (see Chapter 1).

Depending on the physico-chemical conditions, and in particular when the temperature is increased, the dh may also become entropically unfavorable, and the hydrogen bonds between the bases break : this *melting* transition happens at $T \simeq 80$°C at atmospheric pressure and physiological chemical conditions. Here, the influence of the sequence is rather important, because of the different number of hydrogen bonds in the bp, and short oligomers melt also at lower temperatures.

### 0.1.2 Base and base-pair geometrical parameters

In this work, we will be dealing with the deformations of the molecule, *i.e.* deviations wrt the ideal dh described in the previous paragraph. Describing these deformations exhaustively requires to consider the positions of all atoms, which is extremely inconvenient. Luckily, in most cases this can be avoided, thanks to the physico-chemical properties of the organic compounds.

The flat bases are stiff, and exhibit very limited deformations, even under strong constraints : hence, they can be very accurately treated as rigid bodies, as schematically depicted on Fig. 0.1B. This figure shows that the bp are also flat in the ideal dh, however the weaker hydrogen bonds connecting the bases make them much more deformable. The state of deformation of a bp can therefore be described by 6 coordinates, representing the relative translation (*shear, stretch, stagger*) and orientation (*buckle, propeller-twist, opening*) of the two bases : the conventional reference axes are shown on Fig. 0.3A, where each base is represented as a brick.

Assuming that the deformations of the base-pairs remain limited, the next level of description is to consider the base-pairs as rigid-bodies and describe the state of the dh by specifying, again, the relative translations (*shift, slide, rise*) and orientations (*twist, tilt, roll*) of the successive base-pair steps (Fig. 0.3C). This *rigid base-pair* description is very popular for describing oligomers of a few base-pair step (bps), and will also be central to this work, especially because it is the basis for a *mechanical* model of DNA that we will be using (see next section).

The computation of the base and base-pair geometrical parameters from the atomic coor-

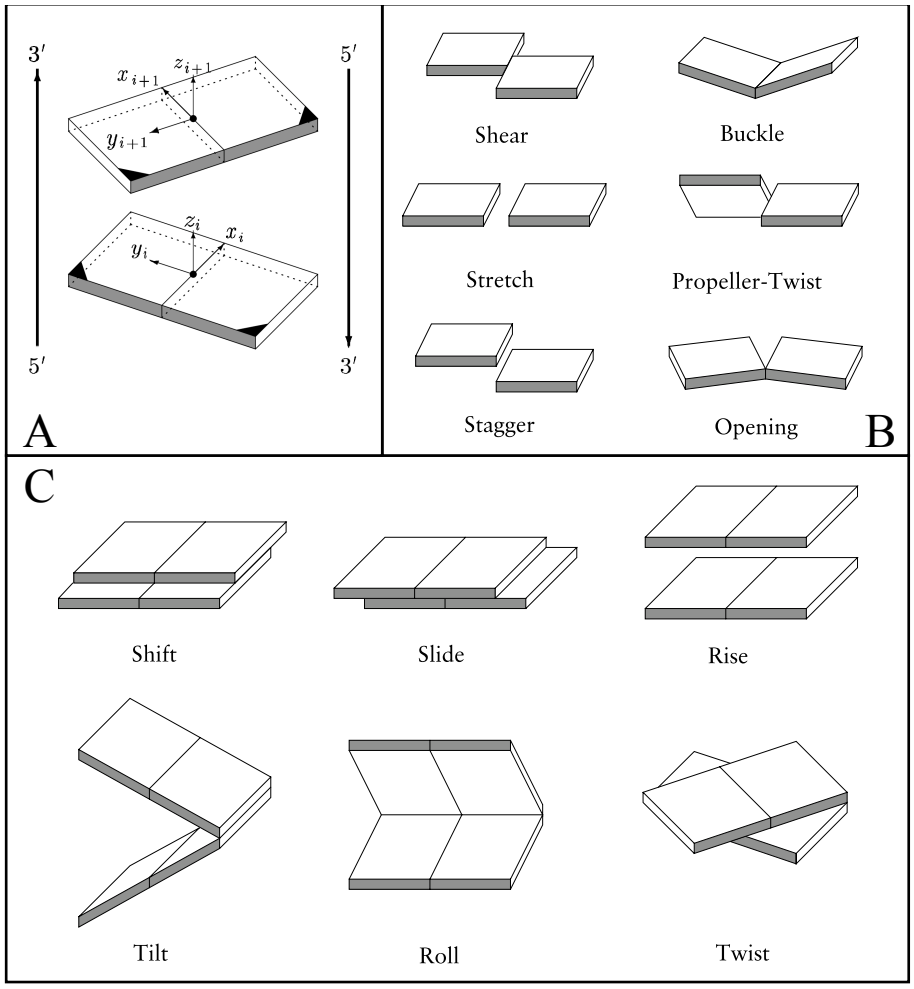


FIGURE 0.3 – Translational and orientational deformations of the base-pairs and base-pair steps. (A) Schematic depiction of the standard coordinate system in the base-pair. The translational and orientational deformations are described with respect to these axis. (B) Internal base-pair deformations. (C) Base-pair step deformations. The bases are represented as bricks, the minor groove side is shaded. Adapted from [Dickerson 1989]

dinates of a DNA oligomer involves the choice of coordinate systems. For the bases, the latter (origin and axes) are defined by the atom positions [Dickerson 1989]. The "Tsukuba" conference was held in 1999 to unite all the community with a single convention [Olson et al. 2001]. For the base-pairs, the choice is far less straightforward because of the internal deformations : and they had been subject to several different choices. Today, small differences subsist in the mathematical definitions of angles, but the following physical choice has been universally accepted : from the center of the bp, the direction $\hat{e}_x$ points in the directions of the major groove, $\hat{e}_y$ points towards the sugar group of the reference backbone, and $\hat{e}_z$ points along the helical axis, in the 3' direction of the reference backbone.

### 0.1.3 Structure of the nucleosome

In eukaryotic cells, DNA is coated with an important amount of proteins, which simultaneously ensures a high compaction level of the long molecule, and allows for local openings

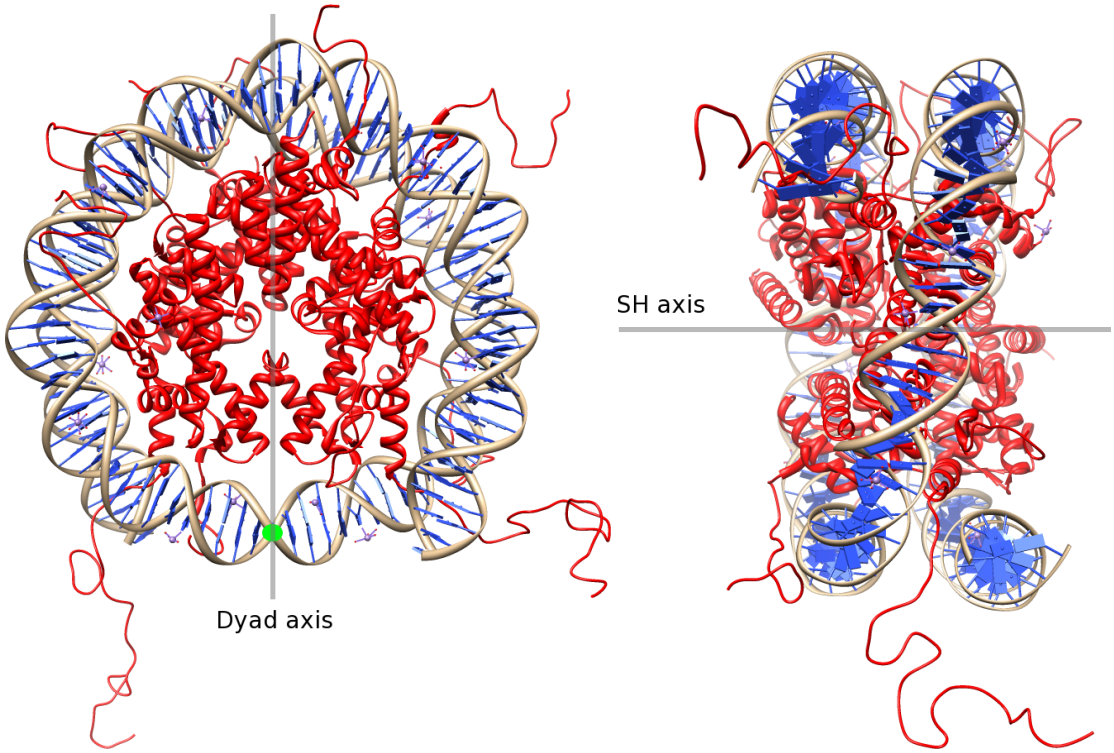


FIGURE 0.4 – Nucleosome Core Particle, in a schematic representation where the histones are colored in red and the DNA in blue (bases) and beige (backbones). View along the superhelical axis (left) and the dyad axis (right), which are shown in gray. The dyad is indicated with a green spot. The flexible tails extend outside the core, here depicted in an unstructured conformation.

required for the expression of genes. This complex is called chromatin, and is schematic depicted on Fig. 1.

The primary constituent of chromatin is the *nucleosome*, which is formed by $200 \pm 40$ bp of DNA, partly wrapped around a complex of 8 histone proteins, containing two copies of each H2A, H2B, H3 and H4. This complex has a roughly cylindrical shape, of 6.5 nm diameter and 6 nm height (Fig. 0.4), which is in close contact with an estimated average of $146 \pm 1$ bp [Richmond and Davey 2003, Makde et al. 2010]. The DNA is wrapped around the proteins in $\sim 1.75$ left-handed helical turns of an approximate *superhelix*, of $\sim 3$ nm pitch. The wrapped DNA and the core octamer form the *Nucleosome Core Particle* (NCP), while the remaining DNA is called the *linker DNA*. The NCP is approximately symmetric wrt *the dyad axis*, which passes through the center of the octamer and the central *dyad* bp located between the entry and exit points of DNA, and is orthogonal to the *superhelical axis*. The superhelical path is indexed by the number of helical turns made by the DNA, with the dyad at the origin : a bp with an integral superhelical location (SHL) are therefore, as the dyad itself, at a point where the major groove faces the octamer.

The histones carry a large number of positive charges, 117 of which are located inside the core. In the physiological conditions where the electrostatic forces are short-ranged ($\sim 1$ nm), only the surface charges are likely to interact with the DNA. 103 further positive charges are located on flexible tails that extend up to 5 nm out of the core, see Fig. 0.4. It is not known if these tails adopt a structured conformation *in vivo*, but they play an important role in chromatin compaction [Dorigo et al. 2003, Shogren-Knaak et al. 2006], most likely by interacting with

neighbor nucleosomes [Arya and Schlick 2009], or with the linker DNA [Angelov et al. 2001]. On the other hand, the core DNA carries an average of $292 \pm 2$, hence a net negative charge for the NCP.

The structure of the core is known at almost atomic resolution from X-ray crystallography experiments [Davey et al. 2002], which are made possible by the rigidity, stability and reproducibility of the complex for some high-affinity sequences. These experiments show that the DNA contacts the octamer in 14 discrete *anchor points*, located every helical turn at the site where the minor groove faces the histone octamer, and where the DNA exhibits severe kinks and shifts. At these points, hydrogen bonds are formed, and positively charged amino-acids (lysines, arginines) come into close contact with the DNA phosphates and even protrude deeply into the minor groove, which suggests that the anchor points are also the main sites of histone-DNA interaction.

The stability of the nucleosome indicates that the gain in free energy due to these electrostatic interactions exceeds the elastic cost of bending the DNA, hence a net favorable free energy. The latter can be estimated experimentally by different methods, which give access to the equilibrium constant of the unwrapped states of the nucleosome. Competitive binding studies [Polach et al. 1995, 1996] measure it indirectly, from the rate of DNA digestion by restriction enzymes, when the site of attack is located inside the core. Fluorescence Resonance Energy Transfer (FRET)-based experiments are more direct : the distance-dependent fluorescence exchange between two dyes is measured with different positions of these dyes on the nucleosome [Li et al. 2005, Tomschik et al. 2005, Koopmans et al. 2009, Gansen et al. 2009], where the signal is different for "open" and "closed" conformations. Altogether, these studies suggest a value of $\sim 1.5 - 2 k_B T$ per anchor point [Schiessel 2003]. This rather limited value is probably essential to the physics of the nucleosome, because unwrapping and other dynamic processes [Kulic and Schiessel 2003b,a] are probably involved in the expression of the genetic material buried inside the core.

And yet, with its average persistence length of $\sim 50$ nm, wrapping DNA around the 6.5 nm diameter octamer has an important energetic cost, which can be estimated to be of $\sim 60 k_B T$ in the Worm-Like-Chain model of DNA (see next section). The crystallographic structures suggest that the cost may even be larger, because of strong localized deformations : for instance, the analysis of the most resolved structural model [Davey et al. 2002] with our parametrization of the rigid base-pair model (see next section) suggests an energy of $500 k_B T$ ! This number must be taken with much circumspection though, because the elastic model is calibrated for small fluctuations around the equilibrium structure of DNA ; for the strong distortions dominating here, the errors are considerable. As a comparison, when the structure is allowed to slightly relax, with maximal displacements corresponding to the atomic position uncertainties in the structure, the elastic energy reduces to $200 k_B T$. Another possible bias comes from the high-affinity sequence used in the crystals. Altogether, we can estimate a mechanical bending cost of 4 to 20 $k_B T$ per site. The mechanical cost and the electrostatic gain are therefore probably in quasi-balance.

### 0.1.4 Nucleosomes and genome accessibility

Bending DNA in a nucleosome imposes strong mechanical constraints, which are more or less easily satisfied depending on the sequence. On the other hand, at the anchor points, the histones are not in direct contact with the bases, which suggests that the histone-DNA interaction

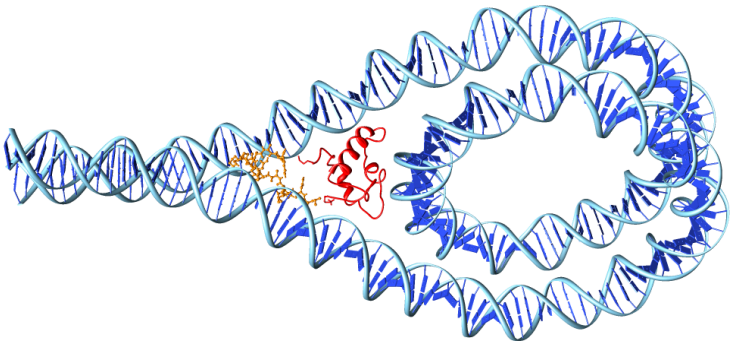

FIGURE 0.5 – Schematic depiction of the stem, with the H1 linker histone shown in red in an arbitrary conformation, with a truncated tail shown in orange. The core histones were not depicted.

free energy is essentially sequence-independent [Olson and Zhurkin 2011]. Therefore, the net association free energy depends on the sequence : some particularly favorable sequences are called *positioning sequences*. They are characterized by the occurrence of repeated favorable tracts at the anchor points, where the deformations are strongest, and thus exhibit characteristic $\sim$10-bp repeats.

Nucleosomes are important actors of genetic regulation. The histones are among the most conserved proteins in eukaryotic world [Nelson et al. 2008], and the mechanisms of chromatin compaction are likely to be very similar among species. In prokaryotes (bacteria, archaea), the genome is much shorter and the mechanisms are different : for instance, there are no nucleosomes, but DNA is still compacted by associating with histone-like proteins [Dixon and Kornberg 1984, White and Bell 2002]. Genome analysis [Ioshikhes et al. 2006, Segal et al. 2006] exhibit a characteristic profile in the gene sequence, where the positioning is strongly favored upstream the gene promoter and disfavored on it.

Indeed, gene expression involves several mechanisms where the presence of the nucleosome core particle is an obstacle. The sequence recognition step may be hindered, if the bases are sterically occluded by the histones [Sahu et al. 2010]. Then, the transcription step by RNA Polymerase involves the opening of the double helix, which dramatically modifies the local state of the molecule and is probably in direct competition with the molecular contacts in the core. This role of the NCP as an obstacle has been verified in various experiments [Li et al. 2007], although not always [Richmond and Davey 2003]. Given the considerable fraction of the genome wrapped in a NCP, this would be problematic if the latter was a static entity *in vivo*.

We mentioned the fact that the two contributions to the nucleosomal free energy were in quasi-balance : the relatively small net free energy allow for spontaneous unwrappings of the ends by the breaking of one or several contact points, and therefore also for the transient exposure of internal sequences to enzymes [Polach et al. 1995]. The transcription of nucleosomal DNA involves a displacement of the molecule by a successive contact-breaking mechanism, where an additional amount of DNA diffuses in the molecule. While this process can already occur spontaneously [Kulic and Schiessel 2003b,a], it is enhanced *in vivo* by active specialized molecules called *chromatin remodelers* [Blosser et al. 2009, Shukla et al. 2010], which are essential to the condensation and decondensation of chromatin during the cellular cycle [Flaus and Owen-Hughes 2001].

### 0.1.5 Linker histone, accessibility regulation and chromatin structure

The accessible character of the genome is therefore crucially related to the possibility of opening and moving the nucleosomes. In the nucleus, this possibility is largely controlled by the

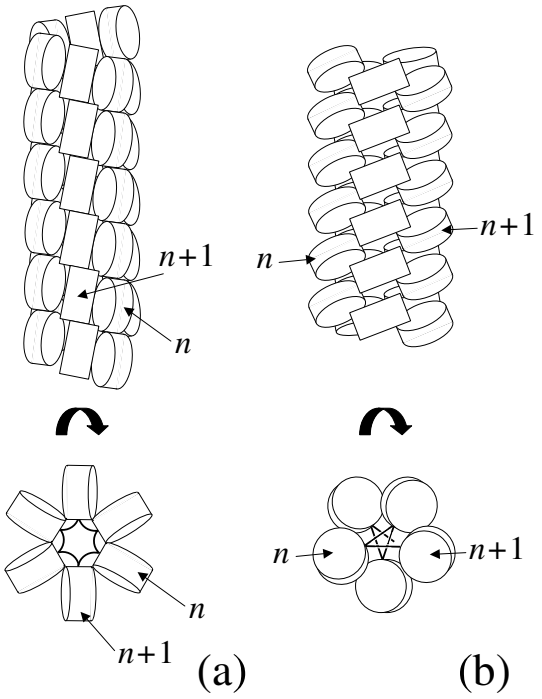


FIGURE 0.6 – Two major proposed models for the chromatin structure : (a) solenoidal model ; (b) crossed-linker model. Source : [Schiessel 2003]

presence of a fifth histone, *the linker histone* H1, which binds DNA close to the entry/exit point. The two branches of the linker DNA come into close contact, forming a *stem* structure [Bednar et al. 1998], which is schematically depicted on Fig. 0.5. This association is made possible by the strong cationic content of the H1 tails, which insert and form structured motifs [Fang et al. 2012] between the two linkers, screening out electrostatic repulsion. In this state, sometimes called *chromatosome*, the mentioned dynamic features of the nucleosome are disabled.

*In vivo,* the stem formation and dissociation mechanisms involve ATP-dependent remodelers and other proteins, which probably induce chemical modifications of the histone tails. The latter modify the repartition of the positive charges, thereby allowing for the insertion or ejection of H1 [Woodcock and Dimitrov 2001, Flaus and Owen-Hughes 2001]. The stem structure is the object of Chapter 3.

The linker histone H1, which is replaced by the structurally similar H5 in some species, is a very abundant protein in the nucleus (around one molecule per nucleosome), and therefore considered as a major actor of gene repression. The repressed state of chromatin, called *heterochromatin*, would thus be mainly composed of H1-bound nucleosomes, and the higher density of this fiber state could be related to the mechanical and electrostatic properties of the stem. Conversely, the more open *euchromatin* would be depleted in H1.

The structure of chromatin results from the folding of the nucleosomal sequence. Two competing models were proposed for this helical arrangement, which differ by the relative position of the successive nucleosomes (see Fig. 0.6) :

- in the *solenoidal* model, they are adjacent, and the strongly bent linker follows the curvature of the core DNA
- in the *crossed-linker* or zig-zag model, they are on opposite sides of the helix, and linked by straight DNA

Most experimental studies have been conducted *in vitro*, and tend to favor the crossed-linker model [Dorigo et al. 2004, Schalch et al. 2005, Tremethick 2007] but some conclude in the

opposite way [Robinson et al. 2006]. These contradictory results may be related to the natural variability of the fiber *in vivo*, that may not be described by a single model. Recent genome-wide studies of molecular factors related to the fiber state, such as the previously mentioned histone and DNA chemical modifications, have isolated a limited set of combinations of these factors, [Filion et al. 2010, Kharchenko et al. 2011], maybe related to a corresponding set of fiber families associated to different levels of genome accessibility. Transition from one state to another could be achieved by active chromatin remodelers.

## 0.2 Physical models

**Physical models for DNA** The models used for describing DNA vary considerably, depending on the length-scale.

- At the smallest scales, a quantum mechanical description is required for computing electronic orbitals and their effect at the atomic scale.
- Next, all-atomic simulations of oligomers use *molecular mechanical* models, with an exhaustive description of the atomic nuclei bound by empirical force-fields.
- Various coarse-grained models have been proposed, where a particle represents a group of atoms. In many cases, the force fields are motivated by structural considerations, and calibrated by computing measurable properties [Mergell et al. 2003, Ouldridge et al. 2010]. In this work, we will use the *rigid base-pair model* of DNA, which is based on the rigid base-pair description presented in the previous section. The deformations of the molecule are treated within the linear response theory, where the elastic parameters can be estimated from the analysis of crystallographic structures or MD trajectories.
- At the larger scale where the bending anisotropy and the sequence effects vanish, the *worm-like chain* models describe the molecule as elastically deformable, with one to four parameters.

In this work, we will use all but the first level of description.

### 0.2.1 Atomistic force-fields

For sufficiently large macromolecules, the details of the electronic orbitals can be neglected. It is then possible to use a coarse-grained description, with empirical force fields between point particles, located at the positions of the nuclei. This *molecular mechanics* description involves a large number of particle types and parameters, in particular because the parameters depend on the prescribed average quantum state of the atoms. In the next paragraph, we describe the models used in the MD simulations of DNA oligomers.

**Bonded interactions** Within DNA, and more generally in macromolecules, the force fields contain bonded interactions between successive atoms, which can be grouped in several categories. Here we describe only the most common ones (see Fig. 0.7). For each category, the parameters of the potential depend on the type of the considered group of atoms : to improve the readability, we do not explicitly write these dependencies.

- *bond stretching* : $v_s = k_s(r-r_0)^2/2$, where $r$ is the distance between two covalently bound atoms. The force constant is of the order of $\sim 800 k_B T/\text{Å}^2$, and they generally deviate only slightly from their rest length. This is the reason why a harmonic function is used in

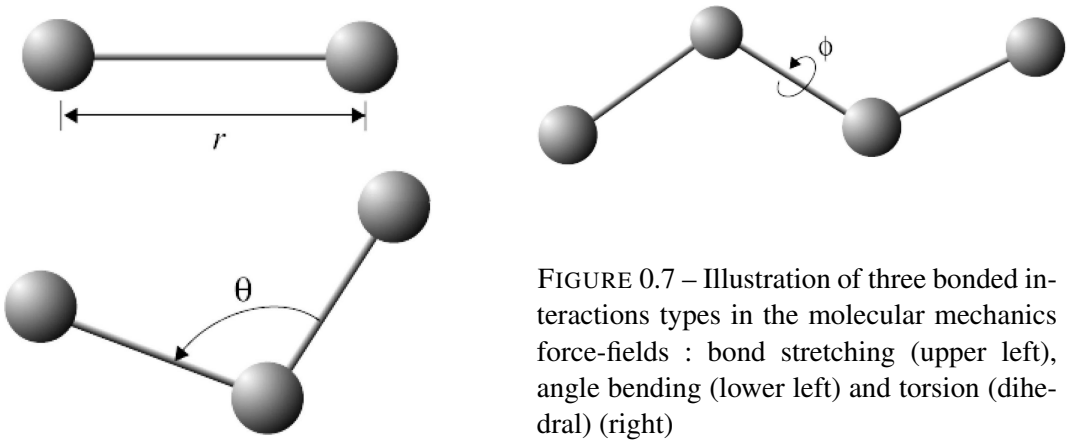


FIGURE 0.7 – Illustration of three bonded interactions types in the molecular mechanics force-fields : bond stretching (upper left), angle bending (lower left) and torsion (dihedral) (right)

place of a more realistic Morse potential, which would require a further parameter [Leach 2001].

- *angle bending* : $v_b = k_b(\theta - \theta_0)^2/2$, where $\theta$ is the angle defined by three successive atoms.
- *torsion* : $v_t = \sum_n V_n(1+\cos(n\phi - \gamma_n))/2$, where $\phi$ is the angle between two planes containing the first and last three atoms in a group of 4 (see Figure). This potential is necessary to reproduce the conformation preferences of organic molecules, in particular the "anti-bonding" repulsive interactions between the two external atoms.

**Nonbonded interactions** The previous potentials apply only to covalently bonded atoms, *i.e.* atoms belonging to the same molecule. The following nonbonded interactions are computed on all pairs of atoms :

- *electrostatic* : $v_q = \frac{q\,q'}{4\pi\varepsilon_0 r}$ where $q$ and $q'$ are the partial charges of the particles and r is the distance.
- *Van der Waals* : modeled by a Lennard-Jones potential : $v_{LJ} = 4E_0\left[(\sigma/r)^{12} - (\sigma/r)^6\right]$ between all particles.

Most force fields include further potentials, such as improper torsion bonds, or dipolar interactions.

**Parametrization of the force fields** With its considerable number of parameters, it is a difficult task to calibrate a given force field. This calibration is achieved by adjusting a given set of experimental results, and increasingly from detailed ab-initio quantum mechanical calculations. The latter are especially used in specialized force-fields like AMBER [Cornell et al. 1995], which is designed for biological macromolecules. The number of particle types is here considerable, because it is a function of the average quantum state of the considered atom, as estimated from quantum mechanical computations of many organic molecular groups. The key hypothesis of the parametrization procedure is *transferability* : the parameters determined for a given chemical group in a molecule can be extrapolated to other (and generally larger) molecules.

**Water model** We will not enter into the details of the parameters in organic molecules, which can be found in [Cornell et al. 1995]. We rather give a short description of a specific sys-

tem : water. Water has been subject to considerable modeling efforts, due to its importance as a solvent, but also because it is both simple and very difficult to model. Since we are generally not interested in the precise position of the water molecules when studying a solute, the best solution would be to describe the solvent as a continuum, which is the basis of the Poisson-Boltzmann approaches and their derivatives. Despite the progress in this line of research [Koehl et al. 2009], the subtle interactions of the solvent with macromolecules often still require to take into account the water molecules and ions explicitly, in which case they consume most of the computing time.

Most water models include 3 or 4 point particles, which are rigidly maintained at prescribed positions. Some more involved models allow for deformations, or include a dipole instead of charges. The different proposed models are parametrized from a combination of thermodynamic quantities, such as the density and the enthalpy of vaporization. In most cases, these calibration points are all at room temperature, and the different models are generally quickly inaccurate when it is varied. The situation is even worse for the phase transitions : for instance, the freezing point of the popular SPC/E model [Beveridge et al. 2004] is 215K at atmospheric pressure [Vega and Abascal 2005] ! Some models were therefore designed for simulations where the temperature varies, such as the TIP4P/Ew model, which reproduces the properties of liquid water over a wide range of temperatures [Horn et al. 2004], and was therefore chosen for our simulations of DNA oligomers at different temperatures in Chapter 1. Note that even there, the freezing point at atmospheric temperature is at 246K... Only a more recent model calibrated simultaneously on properties of liquid and solid water accurately reproduces the experimental temperature [Abascal and Vega 2005].

### 0.2.2 Linear elastic model for a rigid body

In this paragraph, we describe the linear elastic model which will be the central model of this work, applied mainly to the mechanics of a rigid base-pair step, but also to the equilibrium fluctuations of the base-pair internal degrees of freedom and the histone-DNA interactions in the nucleosome. The correspondence between these different applications is given in Table 1.

| Application | Particles | Potential coordinates | Chapter |
|---|---|---|---|
| Rbp model | Adjacent rbp | Watson-Crick bp | 1,2 ,3 |
| Internal bp fluctuations | Rigid bases within a bp | Internal bp parameters | 1 |
| Histone-DNA interactions | Histone core / DNA rbp | Absolute rbp coordinates | 2 |

TABLE 1 – Applications of the elastic model in the following chapters

In all these situations, we describe the state of the system by the relative position and orientation of two interacting particles, *i.e.* a 6-dimensional coordinate vector (q). For simplicity, q will be referred to as the "position" of the conformation. At equilibrium, the system fluctuates around a reference state, with a distribution given by the Maxwell-Boltzmann statistics : $p(q) = p_0 e^{-\beta F(q)}$. Here the Boltzmann weight contains a *free* energy, which refers to the fact that q is the coarse-grained description of an ensemble of microscopic states. The ground state is a minimum of $F(q)$, and in the linear elastic model, the system explores only the deformation states that can be described by the first term in the development of $F$ :

$$F(q) \simeq F(q_0) + \frac{1}{2}(q-q_0)^t \underline{K}(q-q_0) \tag{1}$$

The $\underline{K}$ is a 6x6 matrix describing the resistance of the system to deformation. The ground state $q_0$ is often called equilibrium position, which may be slightly confusing since at equilibrium the mean squared distance to $q_0$ (in translational and orientational coordinates) is nonzero. In this approximation, because the free energy is symmetric wrt the coordinates q, $q_0$ is also the *mean position* of the system.

From the equipartition of energy, it can be shown that the distribution of states is fully described by a $\underline{C}$ :

$$\underline{C} \equiv \left\langle (q-q_0)(q-q_0)^t \right\rangle = (\beta \underline{K})^{-1} \tag{2}$$

**The rigid base-pair model** The most important application of the previous model in this work is *the rigid base-pair model* of DNA. This model relies on several hypothesis :

- the internal bp fluctuations are neglected : a given bps is described by the 6 step coordinates described in the previous section : tilt, roll, twist, shift, slide, rise
- the fluctuations of successive bps are *independent* : the free energy function of a given step depends only on its conformation and sequence
- for a given sequence, the mechanics of the bps is described in the *linear elastic model* developed in the previous paragraph.

An interesting feature of this DNA model is that the parameters can be directly computed from the equilibrium conformational distribution, as expressed in Eq. 2. For a 6-dimensional system, there are 6 parameters for the mean value, and 21 independent parameters for the stiffness matrix, *i.e.* a total of 27 parameters. Here, this number is increased by the sequence-dependence of the elastic properties. When the sequence symmetries of the double-strand are taken into account, there are 10 different base-pair steps, which makes a total of 270 parameters. Note that the model may be "cheaply" extended, by including sequence-dependent parameters at the tetranucleotide level, as suggested by recent MD data [Lavery et al. 2010].

These parameters can be obtained from two sources. The database of crystallographic structures of DNA oligomers or DNA/protein complexes provides an experimental ensemble of conformations, but its interpretation is delicate. Otherwise, trajectories from MD simulations provide a direct estimate of the canonical distribution of states, but they rely on the validity of the molecular models, and are subject to the sampling limitations of all-atom MD. The latter problems and the analysis of MD trajectories will be discussed in detail in Chapter 1. Here, we give some more detail on the former method.

X-ray diffraction of protein or protein/DNA crystals provides detailed information on the molecule. An estimated atomic structure can be inferred from the diffraction pattern by the use of molecular models, with a nearly-atomic resolution. It is with this powerful technique that the dh structure of DNA was resolved in 1953 [Watson and Crick 1953]. The limitations of the method lie, on the one hand, in the necessity to grow a crystal out of the considered molecule, which can only be achieved for relatively rigid molecules. Even then, the specific conformation selected by the process may differ from the conformation (or mean conformation) in solution. On the other hand, the interpretation of complex diffraction patterns also requires the use of molecular mechanical models, which need to be properly calibrated.

In spite of these possible issues, the growing number of reported crystallographic structures of DNA oligomers or DNA/protein complexes provided a unique experimental basis for the parametrization of the rigid base-pair model [Olson et al. 1998]. The hypothesis is that the variety of base-pair conformations in the database is representative of their canonical distribution at some effective temperature. In the case of DNA-protein co-crystals, where the DNA defor-

mations are mainly due to the complexed proteins, this amounts to saying that the sequence is uncorrelated to protein binding, so that a given bps appears in many different deformation states in the database, depending on the protein positions and binding modes : this static randomness is assumed to somehow mimic thermal disorder. The effective temperature must then be estimated, either by fitting the value of coarse-grained observables like the twist persistence length [Matsumoto and Olson 2002, Lankas et al. 2003], or by equating the global fluctuation range with that of a MD canonical ensemble [Becker et al. 2006].

In Chapters 2 and 3, we will use a hybrid parametrization of the rigid base-pair model, where the equilibrium positions are taken from the crystallographic database, while the elastic parameters are extracted from MD trajectories [Lankas et al. 2003]. This combination has proved to be predictive for the binding affinity of proteins by indirect readout, *i.e.* selection of DNA sequences by their facility to accommodate a protein-induced deformation [Becker et al. 2006]. This is consistent with the fact that the force fields involved in the considered MD trajectories were known to underestimate some mean values of the base-pair step parameters, while the mentioned hypothesis of equipartition of energy in the crystals remains hypothetical, and the procedure suffers from a limited sampling.

Recent advances of MD simulations, both in computational power and in the parametrization of the force fields, allowed the "ABC" consortium [Beveridge et al. 2004, Dixit et al. 2005] to sample the sequence dependence beyond the step level : by investigating all tetranucleotides, the results showed that the flexibility depends on the neighbor sequence [Lavery et al. 2010].

**Mathematical description of the elastic model** We have described the physical basis of the rigid base-pair model and its parameters. The implementation of the model used in this work was written as a Mathematica [Wolfram Research 2008] package [Becker et al. 2006], which we extended for specific needs. In this paragraph, we give a short description of the employed mathematical formalism, focused on the practical aspects. More details can be found in [Becker and Everaers 2007, Becker 2007].

In a given reference frame, the position and orientation of a given rigid bp are described unambiguously by a rotation matrix $R$ and a translation vector $p$. The translation vector relates the origin of the frame to the position of the center of the bp, while the rotation matrix transforms the reference trihedron into the local trihedron of the bp. These two quantities can be conveniently combined in a 4x4 "homogeneous" matrix

$$g = \begin{pmatrix} & R & & p \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

If we now consider a rigid base-pair chain, *i.e.* the successive rigid base-pair (rbp) of a DNA oligomer, the *absolute coordinates* of the successive rbp are given by $g_{0,k}$, where the index "0" refers to the laboratory frame, and "$k$" is the index of the rbp in the chain. The advantage of the homogeneous formalism is that, in mathematical terms, the homogeneous matrices form a multiplicative group, which practically means that the matrices can be inverted and combined. The product $g_{k,k+1} = g_{0,k}^{-1} g_{0,k+1}$ is still a homogeneous matrix, which gives the *relative coordinate* of rbp $k+1$ wrt rbp $k$, *i.e.* the coordinates of rbp $k+1$ in the local frame of rbp $k$. In other words, this is the homogeneous matrix representing the coordinates of the base-pair step $(k, k+1)$. These properties make this representation convenient for calculations along the chain.

The next step is to choose a particular set of coordinates for the 6-vector corresponding to a particular matrix $g = (R, p)$. The most common choice is the product of the natural coordinate set for the translation, and a certain choice of Euler angles for the rotation [Lu and Olson 2003], or slightly different choices of angles [Lavery et al. 2009]. Here, we use another coordinate chart, which takes advantage of the fact that the group of g-matrices is a Lie group. Each g-matrix can be generated by the application of a matrix exponential function on a unique combination of infinitesimal generators :

$$g = \exp\ q = \exp(q_i X^i) = \sum_{n=0}^{\infty} \frac{1}{n!} (q_i X^i)^n \tag{4}$$

where $q = \{q_i\}_{i=1,\ldots,6}$ is the coordinate vector and there is an implicit summation on the infinitesimal generators of the group, $X_i$, defined for $1 \leq i \leq 3$ by :

$$X_i = \begin{pmatrix} \varepsilon_i & 0 \\ 0 & 0 \end{pmatrix} \text{, with } (\varepsilon_i)_{jk} = \varepsilon_{jik}\text{, the antisymmetric tensor}$$

$$X_{i+3} = \begin{pmatrix} 0 & d_i \\ 0 & 0 \end{pmatrix} \text{, with } (d_i)_j = \delta_{ij}\text{, the symmetric tensor}$$

The $X_i$ are the respective generators of the rotations around the three reference axis and the translations along them. Although this coordinate chart may seem more abstract than usual decompositions into translational and orientational parts, it is algebraically simpler, since all degrees of freedom are treated in a unique formalism.

For small deformations, the latter expression can be linearized :

$$g = \exp\ q \simeq e + q^i X_i + o(q) \tag{5}$$

where $e$ is the identity g-matrix. The exponential coordinates have a geometrical interpretation : $\omega = \{q_i\}_{i=1,2,3}$ is the angular vector associated to the rotation $R$, *i.e.* it points in the direction of the rotation axis, and $\|\omega\|$ is the rotation angle. The interpretation of the translational part $v = \{q_i\}_{i=4,5,6}$ is less straightforward : it is the initial tangent of the screw motion that joins the identity frame $e$ to the frame $g$ [Becker and Everaers 2007]. In particular, $v$ depends on the rotational part.

It is the latter expansion that will be used to describe the conformational fluctuations of the bps. For a given step of sequence $\sigma$, the mean conformation can be expressed as a g-matrix $g_0(\sigma)$. Then, the conformation $g$ is described as :

$$g = g_0 \exp(q^i X_i) = g_0(e + q^i X_i + o(q)) \tag{6}$$

The linear approximation is justified by the rigidity of double-strand DNA : the width of the angular thermal distributions are below 8°(see p. 48). It breaks in presence of too strong external forces.

**Elastic energy, forces and nanomechanics** So far, we have described the 6-dimensional system in the harmonic approximation of the free energy. Here, we adopt a different point of view, that of an *elastic system* where we can define mechanical forces and elastic energies of deformation.

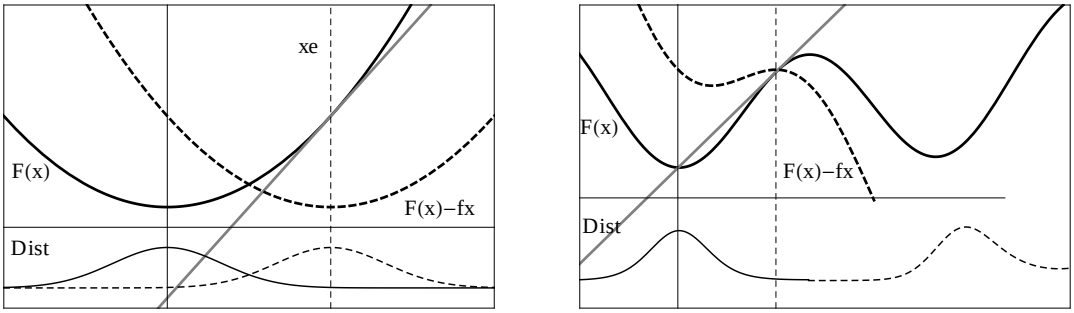


FIGURE 0.8 – Unidimensional examples of energy landscapes, where the approximation of mean force can be applied (left) and crudely brakes down (right). In each case, the upper panel shows the original free energy function (black solid) and the line $fx$ corresponding to the external force (gray solid), tangent to the free energy at the point $x_e$ where the system is constrained. The perturbed function $F(x) - fx$ is the thick dashed line, with its minimum at the constrained point $x_e$. In the case where the approximation is valid (left), the modified distribution (thin dashed, lower panel) is such that the mean position is indeed $x_e$. In this particular case, the original free energy is quadratic, and the modified distribution (dashed) has the same width as the original one. In the case where the anharmonicities are important (right), the construction breaks, the distribution mean being different from $x_e$. Here, the situation is even worse because $x_e$ is in the region where the free energy is not convex, and so $x_e$ cannot be a stable equilibrium.

The free energy function of Eq. 1 is a *potential of mean force* [Chandler 1987] : let us consider a chain of two bp, where the bp 0 is fixed in the laboratory frame, and the bp 1 is constrained in a deformed conformation $q_e \neq q_0$ by an external action, for instance a protein bound to the oligomer, as can be found in crystallographic structures of DNA-protein complexes. Without the protein, the conformation $q_e$ would have a small probability. We now define the *mean generalized force* $\mu_e$, as the force which modifies the distribution of states in such a way that $q_e$ is the new minimum. The approximation is to consider that this new *minimum* is also the *mean position* of the modified distribution : the conditions for its validity are discussed below.

The generalized force is defined as the conjugate variable of the position. The free energy of the step including the external force is given by $F(q) - \mu_e q$, and the condition for $q_e$ being a local minimum is therefore :

$$dF(q_e) = \mu_e(q_e) q_e \tag{7}$$

In the chosen example, the step coordinates coincide with the coordinates of the bp 1 expressed in the laboratory frame, and the force is therefore also in this frame.

This construction does not depend on a particular form of the free energy function $F$ : the condition for the approximation be valid is that the modified distribution is centered at $q_e$. This is granted in particular for a harmonic system, but it may also be true for a system with an irregular free energy function, provided these irregularities are small enough to be averaged out by the thermal distribution. It is not valid if $F$ contains strong anharmonic features in the region close to $q_e$, such as deep secondary minima : in this case, the modified distribution of states could be extremely distorted and the approximation breaks : two extreme cases are illustrated on Fig. 0.8. In our case where $F$ is harmonic (left figure), the approximation is always valid. However, one must keep in mind at this point that the harmonic approximation has been derived from equilibrium fluctuations, and may not describe accurately strongly deformed bps : the limit of applicability is therefore set by the anharmonicities of the free energy, and therefore not known with precision. Several tests operated on crystallographic and NMR structural models of various protein-DNA complexes show that in these range of deformations, the method remains

robust, even wrt rather strong deformations [Becker and Everaers 2009b,a].

In this framework, the system can be seen as a simple elastic body, where the deformed states are described by an elastic energy and exert forces on the connected bp. In the following, the word "force" will always refer to a mean generalized force. The mentioned elastic energy coincides with the free energy function $F$. This usual vocabulary may cause some confusion, because it hides the possible entropic contribution. This is unimportant at a single temperature, but not when the latter is varied, as will be established in Chapter 1.

The deformed bp exerts equal and opposed *internal* force on both connected bp. In this case, the force on bp 1 is given by the negative derivative of the free energy in the laboratory frame :

$$\mu_i(q_e) \equiv -d_q F(q_e) \tag{8}$$

The force balance between $\mu_i$ and $\mu_e$ ensures the static equilibrium of bp 1.

The components of the force, given by Eq. 8, depend on a particular choice of coordinate chart for the 6-vector q. An interesting feature of the exponential chart described in the previous paragraph is that, in the limit of small forces, these components are the conjugate variables of the rotations around and the translations along the three reference axis of the considered bp, *i.e.* the usual torque and force [Becker 2007]. This would not be the case with a different choice of rotational coordinates such as the Euler angles.

An interesting application of this framework is that it allows to computate the external forces exerted by proteins on DNA from high-resolution structures [Becker and Everaers 2009b,a]. The system is now an oligomer, where the bps are deformed by the bound protein. Under the hypothesis that the structure is in mechanical equilibrium, the force balance of each bp $k$ now involves three forces : the external force acting on the bp, $\mu_e^k$, and the two internal forces, exerted by the two steps where the bp is involved : $\mu_i^k = \mu_i^{k-1,k} + \mu_i^{k+1,k}$. Because we know the conformations of the bp and the elastic (free) energy, we can compute the two internal forces easily, for instance by using Eq. 8 in the frame of bp $k$. The external force $\mu_e^k$ is then simply the opposite of $\mu_i^k$.

This *nanomechanical* analysis of high-resolution data allows to explore features that cannot be seen easily in the structures. In particular, the steps may be deformed in a region where there is no force, if the two components of the internal force acting on each bp balance each other. This procedure will be used in Chapter 2 to extract a force field from a set of crystallographic nucleosome structures.

### 0.2.3 From RBP to WLC models : Coarse-graining relations

The RBP model is adapted to the description of DNA oligomers when the sequence effects and the dh geometry must be taken into account. For larger molecules of a few tens nanometers and up, both the sequence specificity and the bending anisotropy fade away. The precise description of the fluctuations of every base or bp becomes irrelevant, because they average out into global fluctuations, which can be efficiently described by a larger-scale coarse-grain model. In the wlc model, the deformations of the molecule are still treated in the elastic approximation, but the number of parameters is considerably reduced. In the simplest version, the molecule is allowed to bend isotropically, with a bending rigidity $k_b$. A more elaborate version takes into account the coupled twist and elongation degrees of freedom, which makes four parameters.

In the simplest description, if we consider the molecule as the discrete succession of rigid monomers of unit size $b$, and direction $\vec{u}_i$, then the bending angle of two successive segments is

$\cos\theta_i \equiv \vec{u}_i.\vec{u_{i+1}}$, and it contributes to the elastic energy of the molecule by a quadratic term :

$$u^{bend}(\theta_i) = \frac{1}{2}k_b\theta_i^2 \tag{9}$$

where we assume that the molecule is straight in average. Within this model, the average angle between distant segments decay exponentially with the distance,

$$\langle \vec{u}_i.\vec{u}_{i+n} \rangle = e^{-nb/l_p} \tag{10}$$

where we have introduced the *persistence length*, defined as the orientational correlation length of the monomers. The persistence length is often used instead of $k_b$ to describe the bending stiffness of the molecule, with the relation :

$$l_p(T) = b\frac{k_b}{k_BT} \tag{11}$$

Computing the value of the persistence length from the parameters of a more detailed model of DNA is not a trivial task. The large-scale flexibility may be influenced by the deformation state of the bp, particularly at high temperature where they become more important (see Chapter 1). However, this contribution is only a correction to the bending between successive bp, and in this work we will restrict to coarse-graining from the step parameters.

The most naive way of doing this relation is an immediate extrapolation from the variance of the tilt and roll angles ($\tau$ and $\rho$ respectively), which are the angles describing the bending state of successive bp wrt the two nonequivalent axis $x$ and $y$ in the bp frame.

For each base-pair, if $\tau$ and $\rho$ remain small, the corresponding rotations commute ; the total bending angle $\theta$ is then given by : $\cos\theta$=$\cos\tau\cos\rho$, and at the first order :

$$\theta^2 = \tau^2 + \rho^2$$

Thus :

$$\langle\theta^2\rangle = \langle\tau^2\rangle + \langle\rho^2\rangle = \frac{k_BT}{k_\tau} + \frac{k_BT}{k_\rho} = \frac{2k_BT}{k_b}$$

where the apparent bending stiffness $k_b$ is given by $k_b = \frac{2k_\tau k_\rho}{k_\tau + k_\rho}$.

From the latter equation and Eq. 11, we finally get :

$$l_p = \frac{2b}{\langle(\tau-\tau_0)^2\rangle + \langle(\rho-\rho_0)^2\rangle} \tag{12}$$

where $\tau$ and $\rho$ are the tilt and roll angles respectively, $\tau_0$ and $\rho_0$ are their mean value, and the monomer size $b$ is the average rise of the considered dinucleotide, $b \simeq 0.34$ nm.

The approximation we made in the previous calculation would be valid only if all bps were parallel and aligned on the helical axis. However, in their actual conformations, they are displaced and rotated from this axis. As a consequence, all degrees of freedom of the steps (and not only the tilt and roll angles) contribute to the large-scale bending. This irregularity is a consequence of thermal fluctuations, but also of the sequence. These effects were taken into account in a rigorous coarse-graining relation developed in [Becker and Everaers 2007], where the irregular helix is mapped into a new rigid base-chain where the equilibrium conformations

are aligned on an axis, and thus the previous calculation can be used. Note that the latter equation provides a "pseudo-persistence length" for each of the 16 dinucleotide sequences, where the prefix "pseudo" refers to the fact that for most sequences, oligomers containing these steps only are forbidden by sequence continuity : they must be understood as the contribution of the considered step to the real persistence length. The sequence-neutral value is obtained from the average of these contributions, for all 16 steps : we will use this computation in Chapter 1. In the case of B-DNA, the deformations of the helix remain limited, and, for the persistence length, the results of the "naive computation" deviate from the rigorous ones by less than 5% [Becker and Everaers 2007].

## 0.3 Simulation methods

Physical models are relevant only when they can be *tested*, *i.e.* compared to experimental results. However, they are generally formulated in terms of abstract quantities that cannot be immediately observed ; the operation generally requires to compute expectation values for some observable. In many cases, this "technical" step turns out to be a major constraint, because only a limited class of models is analytically tractable. In the case of biological systems, it generally requires to wipe off the details of the complex molecular systems. When it is possible, this approach is highly instructive, because it determines generic effects governing systems where the level of detail is often source of some confusion. On the other hand, *each* level of detail is likely to be important at *some* lengthscale of the processes they are involved in. Numerical simulations allow to treat the cases where these details cannot be evacuated. Here, we present shortly two classical simulation methods, which allow to sample the thermodynamic properties of a system at equilibrium.

### 0.3.1 Molecular Dynamics

**Principle** Molecular Dynamics is apparently the most straightforward way to simulate an assembly of particles with a given model for the interactions, such as the molecular mechanics force fields presented in Section 0.2.1. The trajectory is followed for a given total time, by computing the forces and numerically integrating Newton's equations of motion with a prescribed timestep. If the thermodynamic quantities of interest can be expressed as functions of the particle positions and velocities, their value is followed along the trajectory, after an initial equilibration stage. Under the ergodic hypothesis, the mean value of the distribution is equivalent to an ensemble average. Here, by construction the trajectory conserves the total energy, and the sampled ensemble is the microcanonical ensemble. Equipartition of energy suggests to compute an instantaneous temperature,

$$\frac{1}{2}m\sum_{i=1}^{N} v_i^2 = \frac{3}{2}Nk_BT \tag{13}$$

where $N$ is the total number of particles. The time average $\langle T \rangle_t$ provides a measure of the thermodynamic temperature in the system.

**Disorder and determinism** At this point, it is important to notice that equipartition of energy is by no means ensured by construction. Equipartition of energy is a consequence of a disordered

exchange of energy between the different normal modes, resulting in the Maxwell-Boltzmann statistics for their occupancy. This disorder appears as a consequence of "molecular chaos", *i.e.* the random nature of molecular processes. It seems therefore rather surprising that such kind of behavior can be generated by a purely deterministic construction. As an example, it is well-known that a chain of harmonic oscillators has a regular motion, where the normal modes are invisible to each other and do not exchange energy. Simulating such a system in the described scheme accurately reproduces these features if the simulation parameters (integration scheme, timestep) are not too badly chosen, in direct contradiction with the prescribed objective !

The first numerical simulation ever performed [Fermi et al. 1955] was designed to test if molecular chaos is the result of nonlinearities in the system. The answer was positive, but more restrictive than had been foreseen. Only strong nonlinear contributions generated the chaotic behavior : in that case, the limited numerical accuracy results in unpredictable trajectories. On the contrary, weaker nonlinearities result in deterministic (or at least regular) energy flows between modes. This can be a problem, because such regularities are not necessarily apparent in a system with a large number of particles. In our atomistic system, we noted in the previous paragraph that many atomic interactions within the DNA molecule are harmonic, and thus subject to these issues. In this case, equilibration is mediated by the non-bonded interactions, such as the shocks with water molecules, and by the nonlinear bonds when the distortions are severe.

Conversely, if the limited numerical precision of floating-point numbers result in explosively diverging trajectories, it may be surprising that we avoid an accumulation of numerical errors, resulting in ever-increasing energy drifts and finally an explosion of the system ! In fact, such effects can indeed be avoided provided the chosen integration scheme is *symplectic*, *i.e.* preserves the phase-space volume [Frenkel and Smit 2002]. Every timestep, because the Newton equations are typically truncated at a low order of their series expansion, numerical errors deviate the trajectory from the "real" one, but in a way that keeps the system near the same hypersurface of phase space, associated to the total energy E. This deviation is not a problem, since we follow statistical features and not individual trajectories, and we avoid an accumulation of errors resulting in energy drifts. This remarkable property is the main reason why most MD simulations are run with such low-order integration schemes (Verlet, Velocity-Verlet, Leap-frog). Other algorithms approximate the Newton Equations to higher order terms, and may be therefore more precise at short timescales, but they are generally not symplectic, which may result in problematic long-term features [Frenkel and Smit 2002].

**Specific methods** The systems simulated in MD typically contain a large number of particles, and for solvated biomacromolecules, the number of solvent atoms generally exceeds the number of atoms in the solute. For instance, the DNA oligomers simulated in Chapter 1 contain less than 2000 atoms, while there are 45000 solvent atoms in the box ! Most of the simulation time is therefore consumed in the computation of the solvent-solvent interactions, and it is desirable to reduce their number as much as possible. This is in direct contradiction with the objective of sampling thermodynamic quantities, which are strictly valid only for an infinite system. To limit finite-size effects, the most usual trick is to consider *periodic boundary conditions*, where the system is only the repeated unit cell of an infinite lattice. This solution indeed mimics an infinite system, but may also introduce bias and correlations which must be taken into account.

The objective of reducing the computational time lead to several specific methods, which we briefly describe here.

The longest operation at each timestep is the computation of nonbonded interactions, where

theoretically all particles interact, and therefore painfully scales as $N^2$. In fact, it is even worse, since the particles also interact with the infinite periodic copies of the system. Because the magnitude of the interaction decreases with distance, the crudest solution consists in *truncating* the interaction after some distance. However, this is acceptable only for sufficiently short-ranged interactions, where the remaining contribution can be neglected or approximated by a correction term. Here, the efficiency is improved by the definition of "neighbor lists" which store the index of the interacting particles and are regularly updated, so that the distances are not computed every timestep.

For long-ranged interaction on the other hand, such treatments are inaccurate. The typical counterexample is the long-ranged electrostatic interaction, where the interaction energy is not necessarily convergent and the long-range contribution can be important. This interaction is crucial to many macromolecular systems, and a range of techniques was developed to optimize the computation. The general idea of the *Ewald summation* method consists in treating the long-ranged interaction in the reciprocal space, where it is rapidly convergent and can be truncated. *Particle Mesh Ewald* [Darden et al. 1993] is a particular implementation of the latter method, which optimizes the computational efficiency by distributing the charges on a regular grid. In that case, the Fast Fourier Transform can be used, and the computational complexity reduces to $N \log N$.

Another key ingredient for reducing the computation time is to increase the timestep as much as possible, where the limit is fixed by numerical accuracy. For too long timesteps, the integration process may become unstable, which can be easily detected in the simulation because it results in an explosion of the total energy. Because the timestep is essentially fixed by the frequencies of the oscillatory processes present in the system, it is sometimes desirable to freeze the degrees of freedom associated to the highest frequencies. This method can of course only be used if these modes do not interfere with the processes we are trying to simulate. In the case of macromolecules, the vibrational modes associated to the hydrogen require timesteps smaller than 1 fs, which makes them the limiting process, while this degree of freedom can generally be neglected. We therefore introduce modified equations of motion, in which these bonds are *constrained* to a prescribed equilibrium value by Lagrange multipliers. The difficulty in this method is that when an atom is involved in several constrained bonds, the algebraic problem of satisfying them simultaneously becomes difficult. The SHAKE algorithm [Ryckaert et al. 1977] tackles this problem by satisfying each isolated constraint successively. At each step, the procedure may cause another constraint to be violated, and it must be iteratively repeated until all constraints are satisfied within some tolerance [Leach 2001].

**Thermostats and barostats** Finally, we have mentioned that the MD trajectory samples the microcanonical ensemble. However, we are interested at the properties of the system at a given temperature, *i.e.* in the canonical ensemble. While these two samples should be equivalent in the thermodynamic limit, (i) it is not ensured for a finite-size system and (ii) on the practical side, we wish to tune the temperature of the system at a prescribed value. This is achieved by modifying the equations of motion, in a way that mimics the interaction with a *thermostat*. Different schemes were proposed, and we briefly describe advantages and possible issues of the three most common methods for atomistic MD :

- *Berendsen thermostat* [Berendsen et al. 1984] : Eq. 13 immediately suggest a way of adjusting the temperature : compute the instantaneous temperature, and multiply all particle velocities by a global factor that drives it to the desired value, with a certain relaxa-

tion time. This thermostat has the advantage, that it does not strongly perturb the local dynamics. However, there is no evidence that the generated trajectory follows the canonical distribution, where the fluctuations of the instantaneous temperature follow the Maxwell-Boltzmann statistics. If the relaxation is instantaneous, we sample the *isokinetic* ensemble, which is different of the latter by neglecting finite-size temperature fluctuations. For a finite relaxation time, fluctuations remain, but they are generally inaccurate. Recent modifications of this algorithm solve the problem, by computing the global factor from a "goal" temperature picked directly from the accurate distribution [Bussi et al. 2007].

- *Andersen thermostat* [Andersen 1980] :The particles are subject to stochastic collisions with the thermal bath, where their velocity is reset to a random orientation and magnitude, following the prescribed distribution. The advantage of this method is that all parts of the system get equilibrated very quickly, but at the cost of dramatically perturbing its dynamics.
- *Nose-Hoover thermostat* [Nosé 1984, Hoover et al. 1985] and related methods : the method is based on the introduction of a modified Lagrangian containing additional degrees of freedom. It can be analytically shown that for a specific choice of this Lagrangian, an average in the artificially generated ensemble is equivalent to a canonical average of the real system. The advantage of this thermostat is that it relies on exact results. On the other hand, its implementation can be tricky, because the associated dynamics is not Hamiltonian, and more complex versions have been proposed : a discussion on this topic can be found in [Frenkel and Smit 2002]. Also, because the algorithm is deterministic, it is ineffective for thermalizing harmonic systems, in contrast to the Andersen thermostat.

Because numerical simulations have to be compared with experiments, we need yet another transformation to sample the $(N,T,P)$ ensemble, *i.e.* the introduction of a barostat. Here the principle is very similar to the previous one, where the pressure is expressed as a function of the particle coordinates, by the standard definition of the kinetic pressure given by the first term of the virial expansion. The same kind of schemes were adapted to modify the equations of motion, and the same problems may arise.

### 0.3.2 Monte Carlo simulations

Molecular Dynamics simulations sample the (micro)canonical ensemble in a way that closely mimics real trajectories. This is an advantage when one wants to follow the dynamics of the system. However, if only the equilibrium distribution is requested, it may appear as a time-consuming method, in particular because a large number of timesteps can be necessary before reaching a new state decorrelated from the previous one. This is especially true when the energy landscape is difficult to sample, for instance because of high energetic barriers : the trajectory may remain in a single potential well during the whole simulation. In the latter case, simulation techniques exist, that deform the MD trajectory to pass these barriers. Finally, MD is inapplicable to coarse-grained systems where dynamic quantities like the forces are not defined or not easy to compute, for instance when space and time are discretized. In a range of situations, it is therefore easier, faster or necessary to generate a sample of the canonical ensemble from the energy function alone, which is the objective of *Monte Carlo* methods. This name refers to the fact that here, the stochastic character of the distribution is introduced explicitly by random generators.

In this work, we will use only the most simple MC algorithm, where random and independent conformations are generated sequentially, with a probability computed from their Boltzmann weight. We illustrate this method on the very simple example of generating a representative sample of the thermal ensemble, for a rigid base-pair step of given sequence, following the model developed in the previous section. We noted that it is possible to define forces and torques in the coarse-grained coordinates, and therefore to generate this sample by computing a MD trajectory. But how complicated ! Between two independent conformations, several timesteps are necessary, where frame transformations must be applied to compute the forces. Even worse, as explained previously, this harmonic system cannot be simulated without a strongly perturbing thermostat that enforces thermalization. Comparatively, the MC method appears both simple and computationally cheap.

A conformation is generated by choosing the 6-dimensional coordinate from a normal distribution according to the prescribed mean and covariance matrix. This is easily achieved by diagonalizing this matrix : in the base where it is diagonal, one can pick the coordinates from unidimensional normal distributions, where the width is given by the eigenvalue. The chosen conformation is then simply mapped back in the reference base. By construction, all generated conformations are independent, and the computing time is incomparably shorter.

This way of constructing the conformations is only possible in a very simple case where the conformational distribution is known *a priori* by the covariance matrix. In general, the procedure is rather to generate random conformations, where the probability is given by the volume of phase space, compute the corresponding energy, and accept the conformation with a Boltzmann-weighted probability. We implicitly use this scheme for treating volume exclusion : overlapping conformations have an infinite energetic cost, and are therefore always refused.

Simple MC is efficient when the whole phase space, or an important part of it, must be sampled with approximately the same frequency. This is however a rather rare case in condensed matter, where randomly placed particles will almost always overlap, and only a very small subvolume of phase space contributes significantly to the canonical partition function. The conformations generated with the previous method will therefore be nearly always refused, which makes it inefficient. The classical answer to this problem is to generate a *path* in phase space, where the successive conformations are correlated and stay in the same region. In the Metropolis algorithm [Metropolis et al. 1953], this path is generated by successive trial moves, and the acceptance rates are tuned to generate a canonical sample. While this may look like coming back to something resembling MD trajectories, the difference is that all kind trial moves can be chosen. With a smart choice, they will both be frequently accepted, and strongly reorganize the system, in a way that would be extremely long had we followed the physical trajectory : the computing time between two decorrelated points can therefore be enormously reduced. Finally, numerous methods have been developed to bias the selection of conformations, to favor a region of phase space of interest and thereby increase the efficiency when the energy landscape is difficult to sample [Frenkel and Smit 2002].

Note that we may sometimes abusively use the same name “Monte Carlo” in the context of generating samples according to given error estimates, where the error function is assumed to be a Gaussian. The name then refers to the mathematical procedure, which is exactly identical to that of the simple case explained above, even though we do not sample a physical ensemble.

# Chapitre 1

# Entropic contribution in the double-helical elasticity

*In this chapter, we address the question of the temperature dependence of DNA elasticity. This dependence cannot be determined from the analysis of an ensemble of crystallographic structures, because the latter relies on the existence of an effective temperature, which is not related to the real temperature where the experiments were conducted. As a consequence, the temperature dependence of the elastic parameters has been neglected in most DNA models. Here, we conduct all-atomic Molecular Dynamics simulations to estimate the effect of temperature on the structure and flexibility at the base-pair and base-pair step levels. Then, we use coarse-graining relations to relate the computed values to the experimentally measured larger-scale DNA bending persistence length.*

**Conventions and nomenclature :** In this chapter, we will often used reduced units for the temperature : $t = T/T_0$, with $T_0 = 300$K. We express the energies in units of $[k_B T_0] = [k_B].300$K. The lengths are expressed in Å and the angles in degrees. The covariance and stiffness matrices are underlined. For their computation, we use dimensionless units, so that the values of the different matrix elements can be compared (see Section 1.3). The base-pair parameters will be called *intra*-parameters and the base-pair step parameters will be called *step* parameters. Beware that we always plot the persistence length *multiplied by t*, and the covariance *divided by t* : in these representations, a purely enthalpic mechanism yields constant values, and the entropic effect is therefore immediately visible.

## Introduction

DNA is the common substrate of the genomic information of all living organisms - the only exception would be the highly controversial inclusion of RNA viruses in this category [Moreira and López-García 2009]. And yet these organisms live in very different temperatures, especially the so-called "extremophiles", from temperatures below 0°C (in salty water and high pressure) up to 122°C in geothermally heated areas. Most of the latter (hyperthermophiles) are prokaryotes (archaea or bacteria), but some of them are eukaryotes [Zhaxybayeva et al. 2009], the latter being limited to a living temperature not exceeding 60°C [Tansey and Brock 1972]. The discovery of these species was a surprise to biologists, since the physical properties of the

DNA molecule may be rather different in this large temperature interval : in standard conditions, it melts around $\sim$80°C !

As an example of molecular processes where the physics of DNA is essential, let us consider the packaging of DNA into nucleosomes, whose constituents are extremely well-conserved among eukaryotes [Nelson et al. 2008]. The important energetic cost of wrapping DNA around the histones and the gain of interacting with the histones are probably in quasi-balance, and yield a relatively small net free energy [Schiessel 2003], of the order of a few $k_BT$, thereby enabling spontaneous unwrapping [Lowary and Widom 1998] and dynamic features [Kulic and Schiessel 2003b] that are essential to physiological processes. If the mechanical properties of DNA are very different at high temperature, and make it for instance easier to bend while the histone-DNA interaction free energy changes only slightly, then this balance breaks, and the resulting nucleosome will be nastily "glued" together.

This example shows that life at high temperature may face evolutionary challenges at the molecular scale, which explains the biologists' interest in studying the thermophilic species. With the growing database of genomes, many studies tried to relate the living temperature to modifications in the genomic sequence, as a possible adaptation mechanism. In particular, a bias toward more G-C content was expected, because these base-pairs are more stable - principally because they have three hydrogen bonds, while A-T have only two. This question may also be related to the physics of nucleosomes, since the energy required for wrapping DNA in a nucleosome is sequence-dependent. It has been hypothesized that the genome of eukaryotes has evolved to exhibit a "nucleosome code" [Segal et al. 2006], *i.e.* sequence-embedded positioning of nucleosomes, either by favoring certain positions, or by selecting "nucleosome-free regions" that influence the nucleosome positioning in an extended downstream region [Milani et al. 2009, Chevereau et al. 2009]. This thermodynamic landscape is probably not the dominant factor for nucleosome positioning *in vivo*, where active processes may overcome the energy barriers [Zhang et al. 2009], but strong correlations between the sequence and the actual position have been reported in recent experiments [Wang et al. 2008, Brogaard and Widom 2012], suggesting at least a local effect. This positioning is intimately related to the organization of chromatin and the genome accessibility : if the sequence-dependent mechanical properties of DNA change with temperature, thermophilic life may involve modified nucleosome positioning rules.

The results of the genomic studies of thermophiles are contradictory. The expected G-C content bias was not observed [Deckert et al. 1998, Hurst and Merchant 2001, Saunders et al. 2003]. Instead, they appear to be enriched in pyrimidine-pyrimidine dinucleotides [Kawashima et al. 2000]. A proposed explanation is that these dinucleotides are less flexible : the sequence-induced increase in stiffness could somehow compensate the temperature-induced increase of flexibility. Instead of a bias in the sequence, the stability of the double-stranded DNA (dsDNA) in hyperthermophiles could also be sustained by an increase in salt concentration, or by supercoiling, *i.e.* modification of the base stacking by addition or removal of twist [Marguet and Forterre 1994]. The latter hypothesis was supported by the discovery that all hyperthermophiles share a common protein, reverse gyrase, which is precisely able to introduce positive supercoiling in the molecule [Brochier-Armanet et al. 2007, Heine and Chandra 2009].

The accumulating biological data are not yet conclusive ; however they underline the interest of studying the physical properties of DNA at different temperatures. There have been many experimental [SantaLucia 1998] and modeling [Dauxois et al. 1993, Jost and Everaers 2008] studies of the melting transition, and the influence of the sequence and even of neighboring effects [Cuesta-Lopez et al. 2009] on the thermodynamics of the molecule. Much less data is

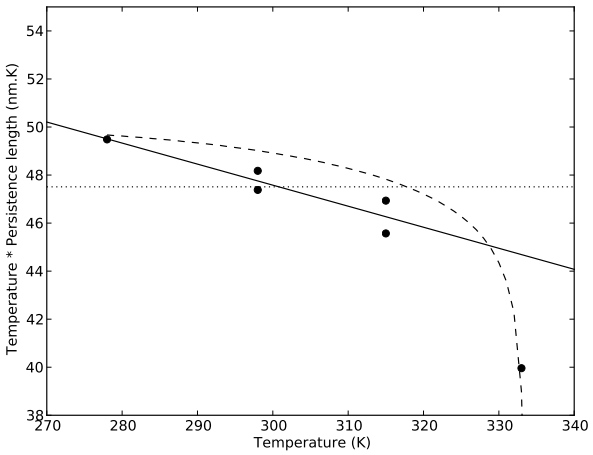


FIGURE 1.1 – Persistence length of DNA measured at different temperatures, between 278K ($t \simeq 0.92$) to 333K ($t \simeq 1.11$). Dots : data from [Geggier et al. 2011]. The last point exhibits a much stronger decrease, which might be attributed to local kinks due to local melting in AT-rich regions. Even if we eliminate this point, the data exhibits a clear entropic contribution to the stiffness, with $t_s = 2.8$ ($T_s = 843K$, solid line). A purely enthalpic model (dotted line) is incompatible with the data. If one assumes that the strong decrease of the last point is due to increased elasticity of the molecule (as the authors suggest) rather than partial melting, the best fit of the data including this point is $T_s = 600K$ (dashed line).

available on its mechanical properties. Recently, the persistence length of DNA has been measured in a range of temperatures between 278K and 335K, for a quasi-random sequence [Geggier et al. 2011], with a combination of cyclization and supercoiling measurements. The $l_p$, defined in Chapter 0, is the orientational correlation length of the polymer : for DNA, it is a measure of the rigidity of the molecule in the large scale where the bending anisotropy vanishes. It can be related to the bending stiffness $k_b(T)$ :

$$l_p(T) = b\frac{k_b(T)}{k_B T} \tag{1.1}$$

where $b$ is the length of a rigid element of the polymer, $k_B$ is the Boltzmann constant and T is the temperature. In the case of a pure bending energy, $k_b(T)$ is independent of the temperature, and the persistence length decreases linearly with $T$. Fig. 1.1 uses a representation ($l_p * t$) where the latter behavior results in a constant (dotted line) : it is clearly invalidated by the experimental data, where both techniques indicate a faster decrease, especially for the higher temperatures (last point).

As an explanation, Theodorakopoulos and Peyrard [Theodorakopoulos and Peyrard 2012], showed in a recent paper that in the considered temperature range, the apparent persistence length of DNA is strongly affected by the appearance of bubbles, *i.e.* locally melted regions of several bps. In these quasi-single-strand (ss) regions, the molecule is considerably more flexible than in the double-helix (dh) state : in effect, they can be approximated by local kinks in the molecule, so that even a small fraction of bubbles can have an important effect on the bending rigidity. The curve (dashed line on Fig. 1.1) shows that this contribution is considerable for temperatures $T > 320\text{K} \simeq 50°\text{C}$, where the persistence length decreases dramatically. For lower temperatures however, the fraction of open DNA is very small, and yields a limited contribution : the decrease on the left part of the curve seems indeed to slow, compared to that of the

experimental points.

One assumption of the model is that the bending stiffness of the dsDNA (closed parts) is independent of $T$, *i.e.* it is treated as pure energy, while the only contribution of entropy to the apparent stiffness is through denaturation. We note however that the dh bending rigidity $k(T)$ is the coarse-grained manifestation of an ensemble of molecular conformations : it is therefore a *free* energy, with an entropic contribution : $k(T) = k_h - Tk_s$ (see the theoretical development in the next section).

In such a model, the enthalpic ($k_h$) and entropic ($k_s$) contributions cannot be determined from an experiment at a single temperature : for instance, the reported value for the persistence length at room temperature is compatible with a whole range of temperature-dependent behaviors (see Fig. 1.2 in the next section). In particular, when $k_s$ is positive, it results in a decrease of the persistence length faster than $1/T$, as can be approximated for a small entropic contribution :

$$l_p(T) = b\frac{k_h - Tk_s}{k_B T} \simeq b\frac{k_h}{k_B(T + k_s/k_h T^2)} \tag{1.2}$$

The solid line on Fig. 1.1 is an example of such behavior, fitted on the data. Here the last point has been excluded from the fit : for the higher temperatures the decrease seems indeed too strong, and we expect the "bubble" contribution to be dominant. We focus on the lower range of temperatures, where the mentioned mechanism may yield a contribution comparable or larger than the bubble formation.

This contribution is also biologically relevant, since it is very unlikely that the DNA is locally melted in molecular processes shared among species living in very different temperatures. As an example, experiments have demonstrated that the relative affinity of different sequences for nucleosome binding is temperature-dependent [Wu and Travers 2005]. While the relatively stiff and spontaneously bent "601" sequence is favored at low temperatures, other more flexible sequences are preferred at higher temperature in the same chemical conditions. This experiment is a direct evidence of the role of entropy in the molecular mechanism of nucleosome binding. Whether this role is mediated by a modification of the DNA elasticity or by other mechanisms remains an open question, but the authors argue for it [Travers et al. 2012].

To quantitatively estimate the entropic contribution to dsDNA elasticity, we have run Molecular Dynamics (MD) simulations of DNA oligomers. This simulation method gives access to the sequence-dependent ensembles of conformations of the dh at the atomic scale, and it has been used to study the sequence-dependent properties of DNA [Lavery et al. 2010]. Here, we extend the study to sequence and temperature dependence. We take advantage of the absence of any local melting in these runs - probably for kinetic reasons - to estimate the temperature evolution of the equilibrium conformations and the flexibility of the dh : our modeling is therefore somehow the counterpart of Theodorakopoulos and Peyrard's model, which included the bubbles but not the temperature dependence of dh stiffness. Because the effects are expected to be small, it is crucial to compute reliable error estimates, and we therefore develop an analysis of the statistical accuracy.

The chapter is divided into the following sections :

1. **Model : Enthalpic and entropic contributions to the elasticity :** We describe a model where the apparent elasticity contains an entropic contribution, which will be the framework of this chapter
2. **Molecular Dynamics of DNA oligomers at different temperatures :** We describe the simulation protocol and show that the distributions of values can be treated in the Gaus-

sian approximation.

3. **Analysis methods :** We develop in detail the methods used for the analysis of the trajectories, in the framework developed in the Model section. *For a first reading, one may skip this technical section and proceed immediately to the results*
4. **Results I :** We apply the analysis procedure developed in the previous section to the distributions of the intra bp parameters. We show that they exhibit a strong temperature dependence, which can be related to the melting path and the spinodal decomposition of the molecule.
5. **Results II :** Here we focus on the step parameters. We find a detectable temperature dependence, mostly of the roll, tilt and rise degrees of freedom. From the computed elastic parameters, we use coarse-graining relations to estimate the temperature-dependent persistence length and discuss the comparison with the experimental datapoints.
6. **Conclusion and outlook :** We discuss the results presented in the previous sections and suggest some possible extensions.
7. **Appendix :** We develop in details the calculations and procedures used in the error analysis of the MD trajectories. Then, we show the detailed plots and results of the sequence-dependent features observed in the MD data

## 1.1 Model : Enthalpic and entropic contributions to the elasticity

### 1.1.1 Unidimensional system

A thermodynamic system at defined $(T,P)$, described by a single degree(s) of freedom (dof) $\theta$, has an elastic behavior if it can be modeled by a free enthalpy :

$$G(\theta,T) = \frac{1}{2}k(T)(\theta-\theta_0(T))^2 \tag{1.3}$$

where $k(T)$ and $\theta_0(T)$ are the stiffness constant and the equilibrium position at temperature $T$. This potential yields a Gaussian distribution of $\theta$ characterized by :

$$\langle\theta\rangle = \theta_0(T)$$

$$\left\langle(\theta-\theta_0(T))^2\right\rangle = \frac{k_BT}{k(T)}$$

In the case where $G$ describes the bending free enthalpy of a polymer (here DNA), $\theta$ is the bend angle, and $k(T)$ is the bending stiffness, which can be related to the persistence length of the molecule :

$$l_p(T) = b\frac{k(T)}{k_BT} \tag{1.4}$$

where $b$ is the length of a rigid element of the polymer.

This is straightforward when considering a single temperature, as in most studies of the DNA persistence length where $T \simeq T_0$.

How does the system behave when T is varied ? For large variations, the system may exhibit features that cannot be predicted by an elastic model, such as DNA melting at $\sim 80°$ Celsius,

where the internal degrees of freedom of the molecule experience a dramatic change. For a small enough temperature interval however, one may consider a linear variation of the free enthalpy with respect to temperature :

$$G(\theta,T) = G(\theta,T_0) + (T-T_0)\frac{\partial G}{\partial T}\Big|_{(\theta,T_0)} \tag{1.5}$$

The second term is the definition of the entropy : $\frac{\partial G}{\partial T}(\theta,T_0) = -S(\theta,T_0)$. Eq. 1.5 can be re-expressed in another way :

$$G(\theta,T) = H(\theta,T_0) - TS(\theta,T_0) \tag{1.6}$$

where $H(\theta,T_0) = G(\theta,T_0) - T_0\frac{\partial G}{\partial T}(\theta,T_0)$ is the enthalpic part of the free enthalpy.

In the linear approximation, $H$ and $S$ are independent of $T$ and yield quadratric contributions to the free enthalpy :

$$H(\theta) = \frac{1}{2}k_h(\theta-\theta_0^h)^2 \tag{1.7}$$

$$S(\theta) = \frac{1}{2}k_s(\theta-\theta_0^s)^2 \tag{1.8}$$

In this simple model, the T-dependent elasticity described by Eq. 1.3 is given by :

$$\begin{cases} k(T) = k_h - Tk_s \\ \theta_0(T) = \frac{1}{k_h - Tk_s}(k_h\theta_0^h - Tk_s\theta_0^s) \end{cases} \tag{1.9}$$

As can be seen, the stiffness constant has a simple linear dependence with $T$, whereas the minimum moves between $\theta_0^h$ and $\theta_0^s$ in a nontrivial way.

It is easier to express these contributions in terms of the stiffness constant at $T_0$, $k_0 = k(T_0)$, and a *spinodal temperature* $T_s = k_h/k_s$, defined as the temperature at which the extrapolated linear $k(T)$ changes sign (Eq. 1.9), and the system becomes unstable. Note that the spinodal temperature $T_s$ is a different notion than the melting temperature $T_m$ of the system. The latter is defined as the temperature where the dh and ss phases of DNA coexist : it depends on the properties of both phases. The dh spinodal temperature on the other hand can be defined from the properties of the dh phase alone, as considered here. The only relation between them is $T_s > T_m$ : only for temperatures $T < T_s$ can the system be in a dh phase (and this is a metastable state if $T > T_m$). As a more familiar example of this phenomenon, while the liquid-vapor coexistence temperature of water at atmospheric pressure is $T_m^{wat} \simeq 100°C$, its liquid spinodal temperature is estimated to be $T_s^{wat} = 330 \pm 2°C$ [Eberhart and II 1985] : liquid water can exist up to $T_S^{wat}$, but not beyond. Eq. 1.9 becomes :

$$\begin{cases} k(T) = \frac{T_s - T}{T_s - T_0}k_0 \\ \theta_0(T) = \frac{T_s\theta_0^h - T\theta_0^s}{T_s - T} \end{cases} \tag{1.10}$$

What can we say of these different constants ? In the temperature interval where the system is studied, the present framework only requires that $k(T) > 0$ : in the opposite case, the system becomes unstable and one has to take higher moments of the free energy into account.

The most straightforward case is the one where entropy contributes to weaken the stiffness of the system, hence $k_s > 0$ and $T_s > T$ where $T$ is a sampled temperature. If $T \ll T_s$, the enthalpy dominates, whereas if $T \sim T_s$, the entropy is important.

Within that case, one sees immediately from Eq. 1.4 and Eq. 1.10 that for a given value of $l_p(T_0)$ (and hence $k(T_0)$), one can consider a whole range of $l_p(T)$ depending on the value of $T_s$. This is illustrated in Fig. 1.2.

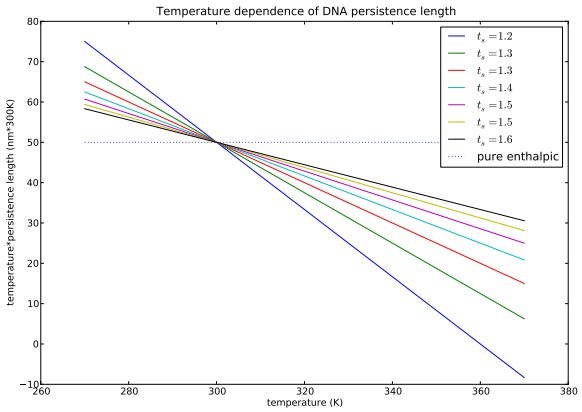


FIGURE 1.2 – Illustration of different evolutions of the persistence length (multiplied by t) with temperature, all compatible with the value $l_p = 50nm$ at $T_0 = 300K$. The solid lines correspond to different values of $T_s$. The horizontal dotted line corresponds to a purely enthalpic system.

On the other hand, for some systems, entropy may have the opposite effect and increases the stiffness. This slightly less intuitive behavior is that of a polymer melt. In that case, $T_s < T$ for $T$ in the interval of study.

### 1.1.2 Multidimensional system

In a $N$-dimensional elastic system, the state is described by a vector $q = (q_1, ..., q_N)$. The physical behavior is similar, and so are the equations. The stiffness is now a $N \times N$ matrix $\underline{K}$, and Eq. 1.3 becomes :

$$G(q,T) = \frac{1}{2}(q - q_0(T))^t \underline{K}(T)(q - q_0(T)) \tag{1.11}$$

In this case, the stiffness constant is related to the covariance matrix $\underline{C}(T) = \langle (q - q_0)(q - q_0)^t \rangle$ :

$$\underline{K}(T) = \frac{1}{k_B T} \underline{C}(T)^{-1} \tag{1.12}$$

This framework is typically used in the rigid base-pair model of DNA, where the base-pair-step (bps) conformation is described by a 6-dimensional vector $q$ =(tilt, roll, twist, shift, slide, rise), the relative orientation and position of the consecutive base-pairs. It can also describe the elasticity of the internal degrees of freedom inside a base-pair in the rigid base model, where the conformation is given by the 6-vector of the relative orientation and position of the two bases : $q$ =(buckle, propeller twist, opening, shear, stretch, stagger).

With definitions similar to Eqs. 1.7 and 1.8,

$$H(\theta) = \frac{1}{2}(q - q_0^h)^t \underline{K_h}(q - q_0^h) \tag{1.13}$$

$$S(\theta) = \frac{1}{2}(q - q_0^s)^t \underline{K_s}(q - q_0^s) \tag{1.14}$$

the equilibrium position is given by :

$$q_0(T) = (\underline{K_h} - T\underline{K_s})^{-1}(\underline{K_h}q_0^h - T\underline{K_s}q_0^s) \tag{1.15}$$

In our study, we fit the values $q_0(T)$ estimated by MD simulations. We therefore prefer a more practical definition, given by the linear development in $q_0(T)$ :

$$\begin{cases} \underline{K}(T) = \underline{K_0} - (T - T_0)\underline{K_s} \\ q_0(T) = q_0^0 - (T - T_0)q_0' \end{cases} \tag{1.16}$$

where $q_0' \equiv q_0'(q_0^h, q_0^s, \underline{K_h}, \underline{K_s})$ is a nontrivial function of the previously defined parameters, which can be more easily determined when fitting the values of $q_0(T)$ observed in the data.

## 1.2 Molecular Dynamics of DNA oligomers

### 1.2.1 Molecular dynamics of DNA oligomers at different temperatures

The effect of the sequence on the DNA flexibility has been successfully investigated by MD for years at the dinucleotide level. Recently, the "ABC" consortium [Beveridge et al. 2004, Dixit et al. 2005] extended the study to all tetramers, which showed that the elastic parameters depend on the neighbor sequence [Lavery et al. 2010]. Some sequences exhibit more complex behaviors, such as bimodal distributions. The extended database also provided elastic parameters for the internal degrees of freedom of the base-pairs, in the approximation of rigid bases.

Here, we focus on the temperature evolution of the DNA mechanics at the base-pair level. We address questions such as :

1. what degrees of freedom, either in the base-pair or base-pair step, are sensitive to temperature changes ?
2. how does this sensitivity depend on the sequence ?
3. can we extract a minimal set of parameters for a temperature-dependent coarse-grained model of DNA, such as the rigid base-pair model ?

Following the ABC protocol, the oligomers used in the simulation were 18-mers built from the repeats of tetranucleotides : the oligomer quoted "xyzw" has the sequence "GCzw xyzw xyzw xyzw GC" (uppercase letters are conserved in all oligomers). For instance, the oligomer AAAC has the sequence GCACAAACAAACAAACGC". To eliminate possible end effects, we excluded the 4 terminal bps on either side of the analysis.

We simulated four 18-mers of dsDNA (AAAC, AGAT, GCGC, GGGG) at 5 different temperatures (273K, 283K, 300K, 325K, 350K) and for 50ns each. The chosen sample contains each of the 10 dinucleotides, appearing as the center of a single tetranucleotide, except for "AA" which appears in two contexts ("AAAC" and "CAAA", both in the oligomer AAAC) : in the latter case, we treated the two tetramers separately, quoted "AA" and "AA1 " (see Table 1.1).

This limited dataset implies that we systematically investigate the sequence-dependence only up to the dinucleotide level. This choice is mainly due to limitations in computing time : with 4 oligomers in 50ns-long dynamics, at five temperatures each, the total trajectory time of 1

| Dinucleotide | Context | Occurrences |
|---|---|---|
| AA | A**AA**C | 2 |
| AA1 | C**AA**A | 3 |
| AC | A**AC**A | 2 |
| AG | T**AG**A | 3 |
| AT | G**AT**A | 2 |
| CA | A**CA**A | 2 |
| CG | G**CG**C | 5 |
| GA | A**GA**T | 2 |
| GC | C**GC**G | 4 |
| GG | G**GG**G | 9 |
| TA | A**TA**G | 2 |

TABLE 1.1 – Dinucleotide sequences

microsecond required the equivalent of 4.5 years of CPU time. As a comparison, to systematically investigate the 136 tetranucleotides requires to run the dynamics of 39 different oligomers, *i.e.* about 10 times more. Moreover, equilibration of the more complex energy landscapes (such as the mentioned bimodal distributions) requires considerably longer trajectories than we did, possibly more than a microsecond at $300K$. The sequence-dependence is therefore not exhaustive, but more detailed than in the previous generation of MD parameters [Lankas et al. 2003].

### 1.2.2 Protocol for MD

The protocol for the Molecular Dynamics simulations was chosen as close as possible to the one used in the ABC study. The AMBER suite of programs [Pearlman et al. 1995] allows to construct each oligomer in the B-DNA conformation, and to run the simulation with a choice of parameters and force fields. The latter are specific to dsDNA, and they have been continuously discussed and improved for the past 15 years, for their ability to reproduce the experimentally observed properties of the molecule, and in particular its crystallographic sequence-dependent conformations. We used the last version, namely the the parmbsc0 modification [Pérez et al. 2007] to the parm99 force field [Cheatham et al. 1999, Case et al. 2005]. The simulations were carried out with periodic boundary conditions, within a truncated octahedral cell. See Chapter 0 for some details on the principle of Molecular Dynamics simulations, and on the atomistic force fields.

Water was modeled using the TIP4P/Ew model, which better reproduces the dynamical properties of water in a broad range of temperatures, as compared to the SPC/E model used in most ABC runs [Horn et al. 2004]. Additional runs with SPC/E solvent at temperatures ($273K < T < 325K$) exhibited values in the same range, but were not analyzed in detail. A typical simulation thus involved around 11 500 water molecules and around 47 000 atoms in total.

Simulations were run with 150 mM KCl, close to physiological concentration, using the parameters from [Dang 1995]. The number of ions was adjusted to ensure a zero net charge for the solute-counter-ion complex. Counter-ions were initially placed at random within the simulation cell, but at least 5 Å from DNA and at least 3.5 Å from one another. The complex was then solvated with a layer of water at least 10 Å thick.

Electrostatic interactions were treated using the particle mesh Ewald method [Darden et al. 1993] with a real-space cutoff of 9 Å and cubic B-spline interpolation onto the charge grid with a spacing of 1 Å (see Chapter 0). Lennard-Jones interactions were truncated at 9 Å and the pair-list was built with a buffer region and list update whenever a particle moved more than 0.5 Å from the previous update.

An initial equilibration stage involved energy minimization of the solvent, then of the solute-solvent system, followed by a slow thermalization, where constraints on the DNA atoms were progressively decreased in 6 successive stages of equilibration-energy minimization, following the standard ABC protocol [Beveridge et al. 2004, Dixit et al. 2005].

Production simulations were carried out in the $(N,P,T)$ ensemble, using the Berendsen algorithm for temperature and pressure [Berendsen et al. 1984], with a coupling constant of 5 picosecond ($10^{-12}$ s) (ps) for both parameters. Note that with this thermostat, the fluctuations of the total kinetic energy may deviate from those of the canonical ensemble.

All chemical bonds involving hydrogen atoms were restrained using SHAKE [Ryckaert et al. 1977], allowing for stable simulations with a 2 femtosecond ($10^{-15}$ s) (fs) time step at all temperatures. Center of mass motion was removed every 5000 steps to keep the solute centered in the simulation cell.

Each oligomer was then simulated for 50 nanosecond ($10^{-9}$ s) (ns) at temperatures 273K, 283K, 300K, 325K, 350K respectively, saving conformational snapshots every 1 ps. This dataset (in a compressed format) requires roughly 430 gigabytes of storage. A second version, without solvent, requires 14 gigabytes.

The atomic coordinates were analyzed by the program Curves+ [Lavery et al. 2009], which computes a full set of helical, backbone and groove conformational parameters. In particular, two sets of parameters describe the state of the base-pair (buckle, propel, opening, shear, stretch, stagger) and the base-pair step (tilt, roll, twist, shift, slide, rise). The definition of these geometrical parameters is not straightforward for a thermally deformed molecule ; they have been subject to several conventions, such as the Cambridge convention for the names and signs of all helical parameters [Dickerson 1989] and the "Tsukuba" reference frame for the description of each base [Olson et al. 2001], which are included in Curves+. Starting from the time series of these parameters as provided by Curves+ (subprogram Canal), the subsequent analysis (Boltzmann inversion, covariance and matrix inversions, error estimates, Monte Carlo (MC) generation of "artificial data", linear regressions, plots) was developed in Python, with the use of the Numpy/Scipy libraries [Jones et al. 2001].

### 1.2.3 Pre-analysis of the data

#### (a) Stability of the simulation and equilibration

The first step of the analysis is to check if the simulated oligomers are stable, and do not exhibit unwanted behaviors (melting, fraying of the ends), in particular for the highest temperature, which is close to the experimental melting point of DNA. Note that the melting point in MD simulations is much higher though, at least in the limited timescale of our simulations. This may be for kinetic reasons, since the melting transition involves passing some energy barriers. On the other hand, the force fields describing the DNA mechanics are calibrated on the structural properties of the dh, and it would not be surprising that the thermodynamics of the melting transition is poorly reproduced. An example of such problems in molecular models

is water, where most models are parametrized from room temperature properties only, and diverge quickly from the observations for higher temperatures ; the chosen model is one of the few exceptions [Horn et al. 2004].

Fig. 1.3 shows the mean values and the standard deviations of the base-pair-step parameters, along the oligomer CGCG : the (odd-indexed) CG steps in the upper panel, and the (even-indexed) GC steps in the lower panel. The values remain in the same order of magnitude all along the oligomer, even at 350K, indicating that the oligomers are indeed stable, and experience no long-acting fraying.

Here the sequence is a succession of alternating CG and GC dinucleotides, and we can compare the values (both mean and standard deviation) at the different positions, as a test for a proper equilibration. The values of the first parameter (tilt, left) are indeed quite reproduced at the different positions of the same dinucleotide for the highest temperatures, as compared to the sequence-dependent variations. For the lowest temperatures however ($T \leq 300K$), this becomes decreasingly true, indicating that the equilibration time is not sufficient for these runs where the kinetics is slower. For some parameters (e.g. shift), the statistical noise for the lowest temperatures is of the same range as the temperature-induced variation of standard deviation in the considered temperature interval. It is therefore crucial to quantify this noise and to take it into account. This is achieved in the next section, by application of the "block averaging" method [Flyvbjerg and Petersen 1989]. Because the insufficiently equilibrated data at the lower temperatures have large error bars, their relative weight in the subsequent modeling is reduced in proportion. To compensate this effect, we have conducted more simulations in the lower range of temperatures, by reducing the temperature interval between two runs.

Apart from this general issue, in two additional runs (at 325K and 350K respectively), some internal-base-pair parameters exhibited values considerably larger than all others, which were interpreted as a sign of premelting and eliminated. The step parameters of the same runs exhibit regular values and were not eliminated.

### (b) Free enthalpy and Gaussian approximation

An elastic model of the base-pair step mechanics approximates the free enthalpy of this step as a quadratic function around an equilibrium value. The validity of this approximation can be tested on the data, by computing the free enthalpy from the distribution of conformations in the MD trajectory. If we note $dq$ a volume element of phase space, we compute the density of states at temperature $T$, $\rho(q,T)$ from the histogram $N(q,T) : \rho(q,T)dq = N(q,T)/N_{tot}(T)$. Then, the free enthalpy $G(q,T)$ is computed as :

$$G(q,T) = -k_B T \log \rho(q,T) \tag{1.17}$$

For this step, we consider that the elementary volume described by the 3 orientational and the 3 translational coordinates is fixed. This is an approximation for curvilinear coordinates, justified by the fact that the angle values remain close to their mean value [Gonzalez and Maddocks 2001].

At this point, the large dimensionality of phase space (6 dimensions) is a major problem : for instance, if we build a histogram with only 10 bins in each dimension, this makes a grid of $10^6$ = 1 million points. For comparison, we have 450 000 data-points for the *most repeated* dinucleotide (GG), and only 100 000 for most others, hence a very poor sampling. This is the reason for the coarseness of the resulting estimation of the free enthalpy, as shown in Fig. 1.4(A).

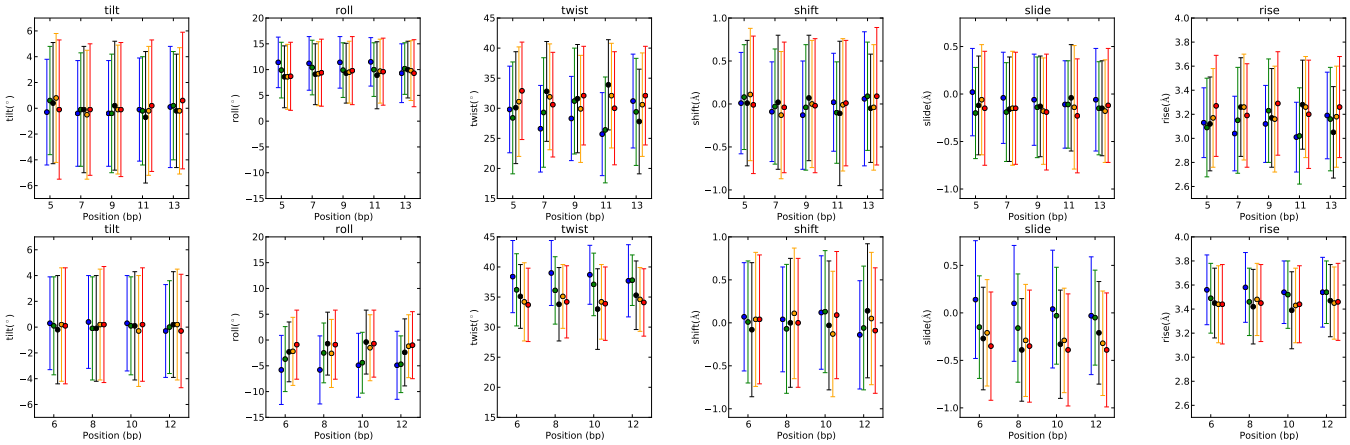


FIGURE 1.3 – Mean values (dots) and standard deviations (bars) of the base-pair step parameters computed along the oligomer CGCG, at the simulated temperatures : 273$K$ (blue), 283$K$ (green), 300$K$ (black), 325$K$ (orange), 350$K$ (red) (these colors will be used throughout the paper). The outer 4 bps on either side of the oligomer have been excluded of the analysis to eliminate possible end effects. **(Upper panel)** Odd bps (CG steps) ; **(lower panel)** even bps (GC steps). The overall regularity of the values show the absence of long-acting fraying or melting, even at $T = 350K$. Comparison of the values obtained at the different positions for the same dinucleotide indicates that for the lowest temperatures, the remaining statistical noise cannot be neglected : for instance, look at the mean values of twist, for the GC steps (lower panel).

To avoid this, we can compute a "pseudo-free enthalpy" by *projecting* the data-points into a subspace of phase space, for instance a given dimension, averaging out all features along the other dimensions. The assumption behind this computation is the independence of the degrees of freedom. Fig. 1.4(B) shows the data projected along the same dimension as shown in (A). Together, these plots show that the free enthalpy can indeed by approximated by quadratic function.

There are however a few exceptions to this statement, corresponding to the bimodal behavior first reported in [Lavery et al. 2010]. Fig. 1.5 shows the estimation of the free enthalpy of dinucleotides "CG" and "CA", projected on the "twist" dimension. It is an interesting question, whether the observed displacement of the equilibrium value corresponds to a structural transition, or just to an artifact due to the poorer equilibration of the runs at the lower temperatures. A comparison of the distributions at the different locations of the dinucleotides shows that the sampling of phase space is indeed not sufficient to resolve this complex energy landscape (see Figure 1.19 in Appendix), which will require a greater computational effort.

In Appendix section (b), we estimate directly the enthalpy and entropy from the free energies computed in this section, and we show that the relative noise prevents the accurate comparison of the curves. We circumvent this problem in the following analysis by computing the Gaussian approximation of the distribution.

### 1.2.4 Covariance and stiffness matrix : reduced units

In the following, we place ourselves in the framework described in Chapter 0 : the distribution is approximated as a Gaussian, and the energy landscape can be described by the estimated second moment of the distribution, *i.e.* the 6x6 covariance matrix $\underline{C}(s,T)$ and its inverse, the stiffness matrix $\underline{K}(s,T)$, where $s$ is the considered sequence.

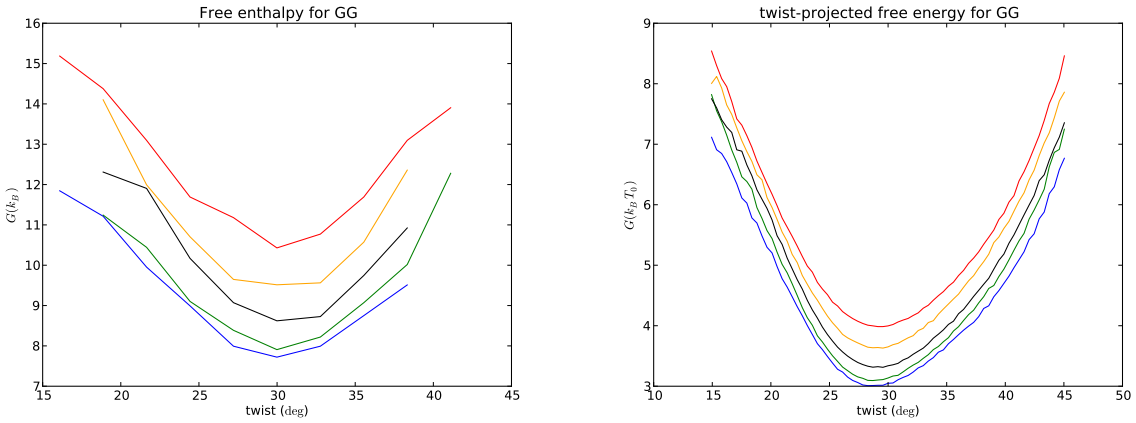


FIGURE 1.4 – (A) Free enthalpy estimation for the GG dinucleotide, plotted on a slice of phase space parallel to the the twist axis, with all other degrees of freedom at their mean value. Temperatures : 273K (blue), 283K (green), 300K (black), 325K (orange), 350K (red). (B) Pseudo-free enthalpy of GG, projected along the twist axis, same colors.

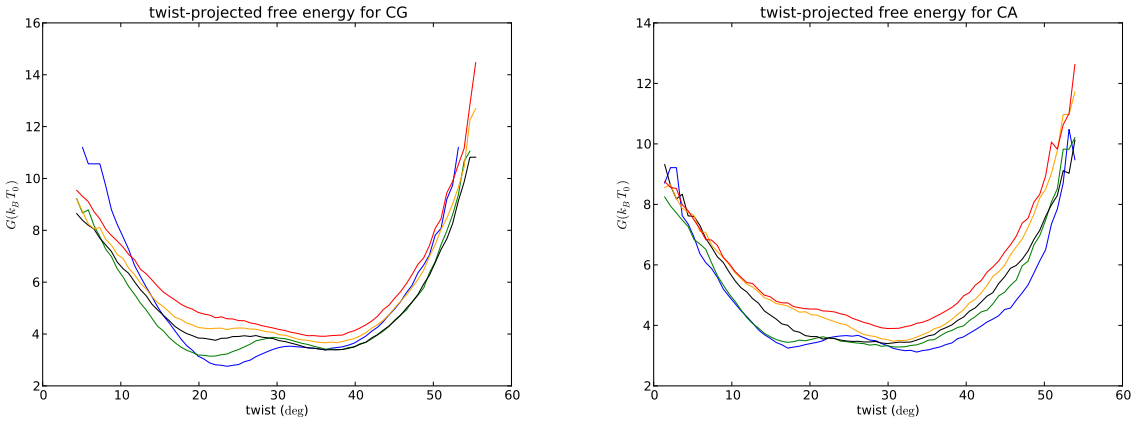


FIGURE 1.5 – (A) Pseudo-free enthalpy projected along the twist axis, for the dinucleotide CG. (B) Same for CA. Phase transition or artifact due to lack of equilibration ?

Because the covariance (and then stiffness) matrices mix the different degrees of freedom, it is convenient to express the values in dimensionless units, so that the different elements can be compared. For each dof, we define the unit $\sigma_i$ as its sequence-averaged standard deviation at $T_0 = 300K$ : $\sigma_i = \sqrt{\langle (q_i - q_i^0)^2 \rangle_{T_0}}$ for $1 \leq i \leq 6$. Hence, in the resulting covariance matrices, the diagonal terms have a value of the order of 1, with sequence- and temperature-induced variations. The average values are given in Table 1.2 for the intra- and step-parameters.

We investigate the effect of the temperature on the elastic properties of the molecule : this involves looking at each element of the 6x6 covariance matrix. As noticed before, it is convenient to look at the *covariance normalized by the temperature*, instead of the simple covariance : in this representation, if the elastic properties are purely enthalpic, the computed quantities are independent of temperature. With the chosen reduced units, the simple covariance is expressed in units of $(\sigma_i \sigma_j)$ and we will use the temperature-normalized units : $(\sigma_i \sigma_j)/[T_0]$.

By inverting the covariance matrix, we obtain the *stiffness matrix*, $\underline{K}(T)$ (see Chapter 0).

| buckle | propel | opening | shear | stretch | stagger |
|---|---|---|---|---|---|
| 11.62° | 9.37° | 4.53° | 0.302 Å | 0.117 Å | 0.425 Å |
| tilt | roll | twist | shift | slide | rise |
| 4.46° | 7.10° | 6.76° | 0.712 Å | 0.712 Å | 0.345 Å |

TABLE 1.2 – Sequence-averaged standard deviations of the different degrees of freedom, at 300K, expressed in degrees (angles) and Ångströms (lengths) : $\sigma_i = \sqrt{\langle (q_i - q_i^0)^2 \rangle_{T_0}}$. All elements of the covariance matrices were computed in dimensionless units, obtained by dividing the data by these reference values. That way, the numerical values of the angle and length diagonal terms as well as off-diagonal elements can be compared. These average values indicate the stiffest degrees of freedom (and hence those with smallest variance) : opening and stretch for intra-parameters, tilt and rise for step-parameters.

The units of the enthalpic and free enthalpic stiffness elements, $k_h$ and $k_g$, are $[k_B T_0]/(\sigma_i \sigma_j)$. For the entropic stiffness elements $k_s$, the unit is the same as the latter divided by the reduced temperature, *i.e.* $[k_B]/(\sigma_i \sigma_j)$.

# 1.3 Analysis methods

In this paragraph, we give a detailed account of the different steps of the data analysis, on the example of the step parameters of the GG dinucleotide. The same method will be used afterwards on the intra- as well as the step-parameters. *For a first reading, the reader may skip this technical section and proceed immediately to the Results.*

## 1.3.1 From time series to covariance and stiffness matrices

The covariance matrix can be estimated directly from the data-points $\{q_i\}_{i=1,\ldots,N}^{s,T}$ of the MD trajectory :

$$\begin{array}{l} q_0(s,T) \equiv \langle q(s,T) \rangle \simeq \frac{1}{N} \sum_{i=1}^{N} q_i \\ \underline{C}(s,T) \equiv \langle (q - q_0(s,T))(q - q_0(s,T))^t \rangle \simeq \frac{1}{N-1} \sum_{i=1}^{N} (q_i - q_0(s,T))(q_i - q_0(s,T))^t \end{array} \quad (1.18)$$

where $q_0(s,T)$ is the equilibrium value. In the estimation of the variance, the sum is normalized by $N-1$ to avoid statistical bias : see Section 1.7.2.

To investigate the possibly small effect of temperature, it is crucial to have an estimate of the statistical noise for these matrix elements. The simplest method would be to compute error bars from the individual values obtained with the considered dinucleotide at its different positions along the oligomer. However, for most sequences, we have only two such positions, which is insufficient. We therefore used a generalization of this simple idea, the *block averaging* method, as developed in the next paragraph.

Fig. 1.6A shows the example of the roll-roll diagonal element of the GG dinucleotide. The data is clearly incompatible with a purely enthalpic model, as can be seen on both covariance and stiffness matrix elements.

As the reader may notice, the analysis of the trajectory of *a single sequence* involves a large body of data : the computation of 6x6 covariance and stiffness matrices for 5 different temperatures, the computation of error bars and a subsequent fit for each of their elements. Showing all this information would be somewhat painful and confusing. Therefore, in this section we show

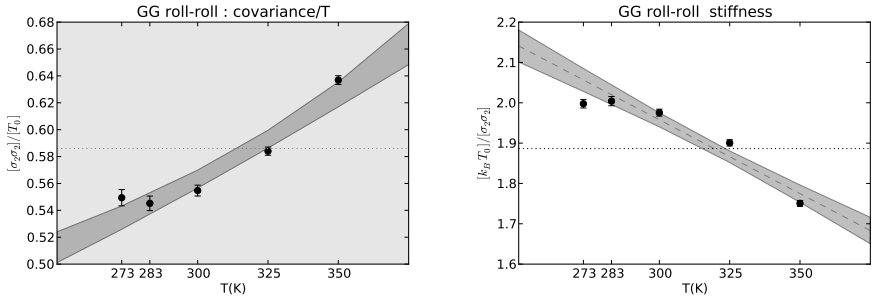


FIGURE 1.6 – Evolution of the roll-roll covariance (left) and stiffness (right) with temperature, for the dinucleotide GG. The covariance is normalized by the temperature, to emphasize the entropic contribution (see text). The data-points (black dots) are incompatible with a purely enthalpic elasticity (dotted line). The error bars on the data-points (here very small) are computed with the "block averaging" procedure (see text). The shaded area is the result of the fitting procedure (see text), with the average estimated value shown as a dashed line for the stiffness. We always plot the covariance on a gray background, to avoid any confusion with the stiffness curves.

only once the whole construction in a single synthetic figure (Fig. 1.8, for the dinucleotide GG), so that the reader may have a global vision of the procedure. When discussing the results of the procedure, we will show only one parameter at a time, as in Fig. 1.6, which is more readable but gives the incorrect impression that the different degrees of freedom can be analyzed separately.

**Error estimates** To estimate the errors on the elements of the covariance and stiffness matrix, as well as the equilibrium values, we used the *block averaging* method [Flyvbjerg and Petersen 1989], which is described in detail in the Appendix, section 1.7.2. The method is a generalization of the simple idea that the typical statistical error can be estimated by comparing the values obtained when splitting the trajectory into smaller parts ("blocks"), and computing the quantity of interest separately on each of them. Instead of choosing the block size - and thus the number of blocks - arbitrarily, we repeat the operation for many different sizes, and estimate the error in each case.

For a normal distribution of variance $\sigma$, a sample of $N_i$ independent points gives the mean value with an expected error of $\sigma/\sqrt{N_i}$. Here, for each block size $n_b$, we compute the mean value on each block separately. The "block error" is then given by standard deviation of the resulting values, divided by the square root of the number of blocks $\sqrt{N/n_b}$. When the block size is small compared to the correlation time of the considered quantity, the block values are not independent, and therefore the computed error is underestimated. On the other hand, for longer trajectories, the samples are decorrelated and the computed error becomes independent of the length : the value of this plateau provides a reliable error estimate.

This behavior is illustrated on Fig. 1.7. The error estimate and a *characteristic time* of the system are obtained by fitting the curve with a simple exponential function (A). One can immediately notice that the quantitative value for the error estimate is not very precise : when the uncorrelated regime is attained, it fluctuates of $\sim 30\%$.

In Appendix Section 1.7.3, we compute analytical results for a simple harmonic oscillator, characterized by a correlation time $\tau$, and we show that the squared error $\sigma^2(m)$ and the correlation time are related by :

$$\sigma^2(m_{T\gg\tau}) \simeq \frac{2\tau}{T}\sigma^2 \tag{1.19}$$

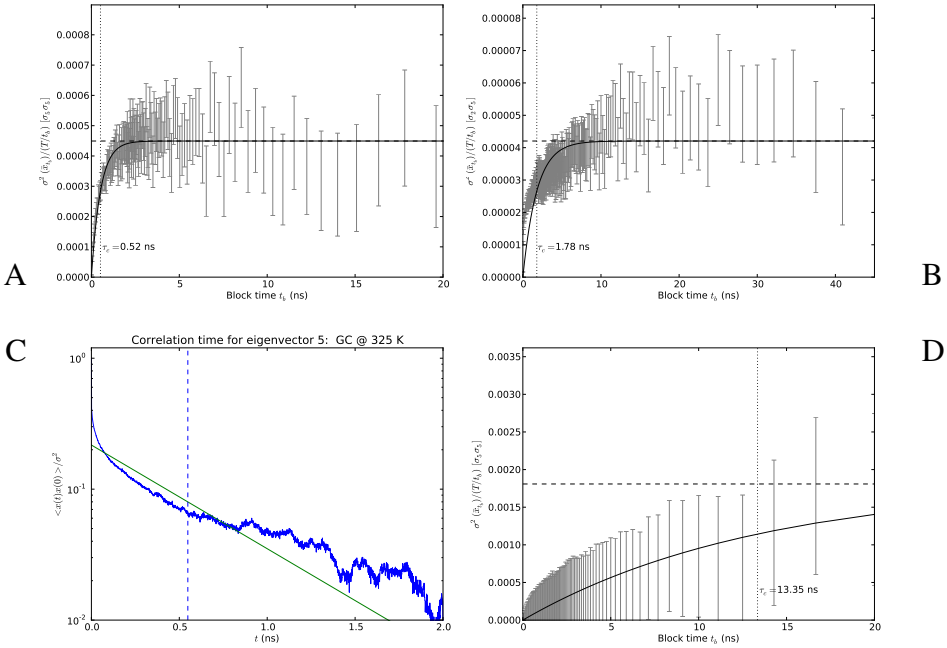


FIGURE 1.7 – Illustration of the block averaging procedure. **(A)** Rise variance of GC at 325K. The "block errors" are estimated with their confidence interval (vertical bars, see Eq. 1.38 in Appendix), and fitted by an exponential function, which gives the estimated error (value of the plateau) and a characteristic time $\tau_c$. **(B)** Same plot for the tilt-rise covariance term : the datapoints exhibit at least two typical timescales : a very short increase and a slower slope before reaching the platau. The fit by a single exponential leads to an underestimation in this case, but within the fluctuations of the computed error. **(C)** Correlation function corresponding to (A), showing that a rapid motion of a few ps dominates the variance, and the fitted timescale $\tau_c \simeq 0.5$ns corresponds to the slow motion fitted in the block curve. In this case the error is dominated by the rapid motion. **(D)** Rise-rise covariance of AA at 300K : the data exhibits a very slow motion, and our sampling is insufficient. In this case, we extrapolate the error points to an estimated constant regime, and we multiply the error range by 2 for security.

where $T$ is the total trajectory time and $\sigma^2$ is the variance. Note that this is just the continuous version of the result used to compute the block averages, and it allows to compute the correlation time of the data from the fitted error (the estimated value of the plateau). In a dataset containing a single correlation time, we also unsurprisingly show that in the "block curve", the fitted characteristic time is the correlation time (with a factor 1/2). This relation is not observed in the MD data : the fitted characteristic times are typically of the order of $0.5 - 2$ns as in (A), while the correlation time computed from the fitted error is typically in the range $50 - 100$ps, *i.e.* 10 times less !

To better understand if this surprising observation results from an error in the procedure or is a feature of the data, we notice that the hypothesis of a single correlation time is not adapted for our trajectories. First, the system is 6-dimensional, with different levels of stiffness associated to the different dof. Because the correlation time of an elastic system depends on the stiffness constant, the different times associated to the different degrees of freedom are likely to be different from one another. These dof are correlated, and we therefore expect the covariance elements to contain a mixture of these times. Therefore it is not surprising that the "block curves" deviate from a simple exponential function, as in (B) where the presence of

several timescales is visible.

In Appendix Sec. 1.7.3 we compute the properties of a system composed of the superposition of two independent harmonic oscillators, with different correlation times. In analogy with Eq. 1.19, the error can then be related to the *total correlation time*, but in this case the latter is different from the *characteristic time* visible in the plot and estimated by the fit, which is essentially the slowest time of the two processes.

Interestingly, the mentioned surprising relation between the two computed times can be rationalized if the analyzed process is the sum of a very rapid motion $x$ that explores large values, and a slower motion $y$ that explores a smaller range of values : $\tau_x \ll \tau_y$, $\sigma_x \gg \sigma_y$. This behavior can be understood as a fast local motion in the base-pair step (bps) that explores most values, and a slower motion associated to large-scale rearrangements of the oligomer, which perturbs the local conformations only little. In that case, the total correlation time is given by (Eq. 1.62) $\tau_{tot} \equiv \tau_x + \frac{\sigma_y^2}{\sigma_x^2}\tau_y \ll \tau_y$, while the decay visible in the curve has a characteristic time of the order of $\tau_y$ : in this specific case, the error is therefore not governed by the slowest time, but rather by the shortest ! To validate this explanation, we plot in (C) the correlation function of the same degree of freedom (rise) sampled in (A). The curve shows that after a few ps, the correlation function has already fallen to 0.3, which indicates that the large variations are indeed dominated by a rapid motion. The curve then exhibits a slower decay, with a correlation time of $\sim 0.55$ns, compatible with the slow motion visible in the block curve. In our procedure, the main limitation introduced by the slower time is that when the "block curve" attains the plateau regime, the statistical uncertainties are already considerable, and the error bar cannot be estimated with more than $\sim 30\%$ precision. In other words, the slow motion cannot be sampled with sufficient accuracy, and its contribution may be underestimated.

Finally, in some rare cases the characteristic time is so large (or conversely, our sampling time is so short !) that we never reach the plateau (D). In that case, all block sizes underestimate the error, and the fitting procedure extrapolates the curve to find the constant ; for security, we increased this value by an growing factor (between 1 and 2), as a function of the fitted timescale of the curve.

### 1.3.2 Linear model of the stiffness temperature dependence

In the framework of the model developed in Section 1.1, we fit the stiffness matrix linearly :

$$\underline{K}(T) = \underline{K_h} - t\underline{K_s} = \underline{K_g^0} - (t-1)\underline{K_s} \tag{1.20}$$

where $\underline{K_h}$, $\underline{K_s}$ are the enthalpic and entropic contributions to the stiffness, and $\underline{K_g^0}$ is the stiffness at $T_0 = 300K$.

**Fitting procedure** Modeling of the data according to Eq. 1.20 implies a linear regression of the data-points for each element of the matrix, where the slope of the fit is the opposite of the entropic contribution to the stiffness element, $-k_S^{ij}$, and the displacement at $T_0 = 300$K is $k_g^{0,ij}$.

We fitted independently all elements of the matrix, with the weighted fitting procedure from [Press et al. 2007]. The procedure consists in minimizing the square deviation of the observed values, as compared to the fit, weighted by the inverse of the uncertainty. The latter choice reduces the weight of the lower temperature runs where the equilibration is poorer, and thus the error bars larger. The advantage of this fitting procedure is that it can be solved analytically, and

yields not only the parameters of the fit, but also error estimates for the enthalpic and entropic stiffnesses, as well as correlations between them : together, these quantities allow to estimate the values and errors of $k_h = k_g^0 + k_s$ and $t_s = 1 + k_g^0/k_s$, and more generally the value and the error for the stiffness matrix at any temperature. In the plots, we therefore depict the model not as a single line, but rather as a shaded area. The concave shape of the area is the result of the correlation between errors in both coefficients of the fit.

Fig. 1.8 shows the complete fitting curves of dinucleotide GG. The upper (white background) triangle shows the fits of the stiffness matrix elements and the lower (gray background) triangle is the corresponding covariance matrix.

One must note that this procedure has some limits, which affect mainly the cases where the sampling and the effect of temperature is limited, in particular the step parameters at the lowest temperatures. In these cases, the fitted data exhibits not only larger error bars, but in some cases the computed value for one or several temperatures do not fit with the other values, in a way that exceeds the estimated statistical error (*e.g.* roll-tilt at 283K). An explanation for such systematic errors could be the existence of metastable states ; this hypothesis seems compatible with the observation that these shifts affect several elements of the matrix simultaneously. Such metastable states could be similar to the bimodal distributions observed in ABC [Lavery et al. 2010]. Longer trajectories indicated that in some cases, these distributions are artifacts due to the limited sampling, but in other cases they seem to indicate actual features of the molecule [R. Lavery, private communication]. For off-diagonal elements, the error bars are sometimes clearly underestimated (for instance for the rise-twist element), but these cases are also those where the variations of the values are near the noise level. In the previous section, we mentioned that the error range is estimated with poor accuracy : in this case, this effect is visible.

Inconsistent points, or "outliers", are not well-treated by a simple least square procedure and we tried to improve this step by using a robust fit [Press et al. 2007] which reduces their weight by modifying the error function employed in the minimization procedure. The result was indeed better when only one point was erroneous, but because of the limited number of datapoints (5 for each fit), the improvement was not considerable when the data was very noisy. Importantly, in the robust fitting procedure, the minimization of the error function has to be computed numerically, and does not provide confidence intervals for the best fit parameters. We implemented a procedure that computes these errors numerically, by changing artificially the position of the fitted data-points, but even then the improvement was not significant. Because these problems occur mainly in the cases where the effect of temperature is not important, they have little consequence on the overall results, even if they are apparent in some specific plots. We therefore keep using the standard least-squares procedure.

**Testing the temperature dependence** For some elements of the matrix (for instance, the roll-roll term (2,2)), the data clearly exhibits a temperature dependence, confirming that it has the statistical quality to resolve the effect we are investigating. On the other hand, for other terms (for instance, the twist-rise coupling term (3,6)), no clear tendency emerges out of our data. In the former case, an accurate value of $t_s$ can be computed, characteristic of the relative weight of the entropic contribution.

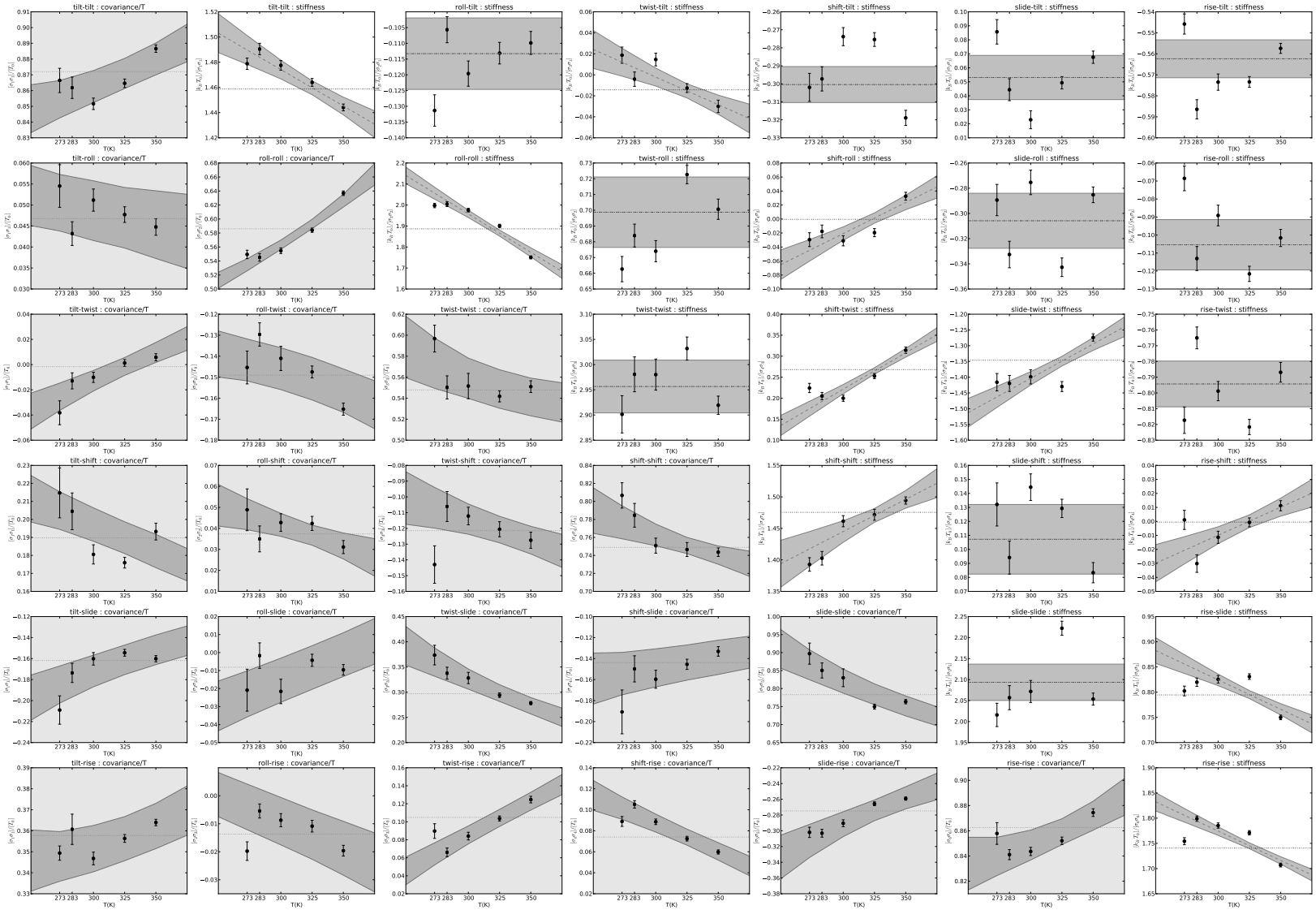


FIGURE 1.8 – Synthetic illustration of the complete fitting procedure of the GG step parameters. The elements of the stiffness matrix are shown in white background : datapoints with estimated error and fitted model (shaded area). The covariance elements are shown with gray background : because both matrices are symmetric, the corresponding elements are related by a transposition operation. The datapoints are shown with their block-estimated error. The model is obtained by fitting the stiffness points, and the result is inverted to compute the covariance model. In the latter opertation, we compute the error bars Monte Carlo, assuming that the errors of all stiffness elements are independent : this results in slightly overestimating the error on the covariance. This effect compensates for the underestimation in the error estimates, which is due to the limited sampling.

The coexistence of these two situations is not surprising. Indeed, the cross-terms couple different degrees of freedom, characterized by different stiffnesses and responding differently to the temperature increase. With the limited sampling, (i) a certain level of noise in the data may dominate the variations of the degrees of freedom less sensitive to temperature, and (ii) the mixing of different degrees of freedom with different characteristic times may result in hiding some of the variations. We therefore introduce a quantitative criterion, to discriminate the degrees of freedom where the effect of temperature can be quantified, from those where it is too small to be detected in the data.

The "f-test" is a statistical test on data for which alternative models exist, of different numbers of parameters : here, the purely enthalpic model where the stiffness is a constant (1 parameter) compared to the 2-parameters (enthalpic and entropic) model. The latter fit will always be more precise than the former, but how much ? The test gives a measure, if the gain in precision justifies the introduction of a new parameter [Schumacker and Lomax 2004].

The criterion for accepting the regression is a threshold on a number, computed from a combination of this test, with a contribution of the relative uncertainty on the temperature $t_s$ and of Pearson's correlation coefficient between the data-points. The two latter contributions were added to eliminate some specific cases, where the f-test was positive while the results of the fit were not acceptable, in particular because the uncertainty on the slope was considerable, and hence the temperature $t_s$ could not be defined. In most cases however, all three contributions go in the same direction, and the overall results are not very sensitive to the precise value chosen for the threshold.

**Matrix inversion and error estimates** The linear fit is realized on the elements of the *stiffness matrix*, obtained by inverting the covariance matrix computed from the datapoints. In the inversion operation, the different elements of the matrix get mixed, and the consequences of a certain level of noise on the computed quantities is therefore not straightforward. To validate our fitted model, we therefore prefer to compare the *covariance* elements computed directly from the trajectories, with a predicted covariance matrix obtained by inverting the fitted stiffness model, including the error estimates. In Fig. 1.8, this step means passing from the shaded areas on the upper triangle of the matrix (stiffnesses, white background) to those of the lower ones (covariances, gray background).

For this operation, we made the simplistic hypothesis that the errors of the different stiffness elements are independent. In that case, from the parameters and errors of the fitted *stiffness* model, one can simply generate a set of stiffness matrices of appropriate statistics by MC, invert each of these matrices, and estimate the error on the *covariance* from the standard deviation of the distribution. This procedure is consistent with the independent fitting of the different stiffness matrix elements ; in the real data however, they are correlated, and the proper treatment of these correlations would imply the computation of their cross-correlations, *i.e.* the computation of the fourth moment of the distribution, a 21x21 matrix of 231 elements ! Because of the limited resolving power of the sampling, we limited our analysis to the second moment. As a result, the computed error bars for the covariance may be overestimated.

The inspection of the different covariance elements of Fig. 1.8 shows that the uncertainties from the model are indeed larger than the error bars estimated directly on the data. On the other hand, we noticed in the previous paragraphs that the error bars may be sometimes underestimated, and this effect somehow compensates for that problem. Most importantly, the temperature dependence still clearly emerges out of the noise. Notice that as a consequence of the inversion

operation, the covariance model curves are all at least slightly temperature-dependent, even when the corresponding stiffness element is modeled as a constant.

To get a further insight into the effects and problems of each step of the error analysis, we have generated artificial trajectories that mimic the properties of the real data. The advantage here is that we know the "exact" model by construction : when applying our procedure, one can then compare the estimated errors with the real ones. The whole procedure is described in Appendix Section 1.7.4 : we show that (i) the estimated error bars are indeed often underestimated, in the range $\sim 30\%$ corresponding to the uncertainties introduced by the slowest motions ; (ii) after inverting the fitted model, the resulting error bars are indeed larger than those estimated directly on the covariance matrix, as already noticed on Fig. 1.8. They are also generally of the same order of magnitude or slightly larger than the real error, and can therefore be considered acceptable.

### 1.3.3 Sequence-dependence, normality and parameter set

It is an interesting question, whether the entropic contribution to the stiffness is similar among the different sequences (even though the stiffnesses themselves are different), or if they are sequence-dependent. In the former case, the parametrization of a T-dependent elastic model could be achieved with much fewer additional parameters than previously mentioned.

It is therefore tempting to operate the previous method on a sequence-averaged dataset. However, the condition for this step is that the resulting distribution remains Gaussian ; this is true only if the standard deviations of the different sequences are similar, and if their mean values are close (as compared to the typical standard deviation) : in other words, if the elasticity itself is sequence-independent. If it is not the case, the covariance computed on the mixed dataset is an artifact of the construction rather than any physical feature of the system.

Some bp parameters do exhibit a very limited sequence-dependence, and can be treated together, either for the group of all sequences, or by grouping the sequences where the considered bp is the same (either A-T or C-G). It is never the case for the step parameters in our dataset, even for the two different 'AA' dinucleotides found in two different tetramers, as illustrated on Fig. 1.9 : their respective covariance ellipses clearly divide into 2 groups (corresponding to the two sequences). Computing the covariance on the aggregated data would yield unphysically large variances.

In some cases, we have tested if the number of parameters can be reduced, by considering a common value of $t_s$ for a group of sequences which have different estimated $k_g^0$ and $k_s$. The method is the following. We first operate the described method separately on all sequence-specific distributions, *i.e.* on distributions where the Gaussian approximation is acceptable. Even though the stiffnesses at room temperature $k_0$ are rather different, for some degrees of freedom we find that the relative weight of the entropic contribution (measured by the parameter $t_s$) is indeed in the same range for all sequences, or for a group of sequences. In that case, we estimate the best common value of $t_s$ for the group, while allowing different values of the stiffnesses $k_g^0$. For each sequence, we divide the data by the sequence-specific value of $k_g^0$ : that way, the data from the different sequences fall together and can be fit as a single dataset, yielding a global $t_s$ and a global $k^0_{g,glob}$ (of the order of, but possibly not equal to 1). Multiplying back by the previously found sequence-specific values of $k_g^0$ keeps $t_s$ unchanged for all sequences, and the corrected (sequence-specific) value of $k_g^0$ is $k_g^0 k^0_{g,glob}$.

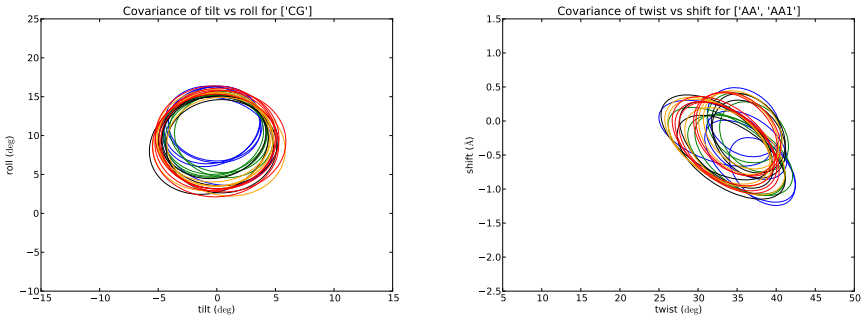


FIGURE 1.9 – Covariance ellipses computed on the individual positions for the different temperatures (from blue-cold to red-hot). (Left) Rise-slide coupling for the CG dinucleotides. Despite a remaining level of statistical noise, one can follow the evolution of the covariance with temperature. (Right) Twist-shift coupling for AA and AA1 dinucleotides. They clearly divide into 2 groups (corresponding to the two sequences), especially for larger temperatures where the system is better equilibrated, confirming the influence of the neighbor bps on the dinucleotide conformations.

We then choose between the most detailed set of parameters (*i.e.* different values of $k_0$ and $t_s$) and the more "economic" one (*i.e.* different values of $k_0$ and same $t_s$), once again, by the use of the "f-test" that compares their predictive power.

### 1.3.4 Equilibrium values

So far, we have considered only the covariances, *i.e.* the second moments of the distributions. For the equilibrium positions, we fit the data at the different temperatures in a very similar way, by using Eq. 1.16 :

$$q_0(T) = q_0^0 - (T - T_0)q_0' \tag{1.21}$$

where $q_0^0$ is the equilibrium conformation at $T_0$ and $q_0'$ is the slope of the fit. We then evaluate the predictive power of this fit, as compared to a temperature-independent value. This test determines the degrees of freedom where entropic effects modify the equilibrium position. The computation of error bars follows the same guideline as for the variance.

## 1.4 Results I : Internal base-pair parameters

In the next paragraphs, we apply the analysis method developed in detail previously, in the case of the internal degrees of freedom of a base-pair. The framework is the model described in Section 1.1 : we describe the system as an elastic medium, where the stiffness is composed of an enthalpic and an entropic contribution. From Eq. 1.16, for each element of the stiffness matrix one can write :

$$k(t) = k_0 - (t - t_0)k_s = \frac{t_s - t}{t_s - t_0}k_0 \tag{1.22}$$

where we have defined *spinodal temperature* $t_s$, at which the considered element of the matrix changes sign, and thus the system becomes unstable in the associated direction. See Section 1.1 for a discussion on its meaning, and the difference with the *melting* temperature $t_m < t_s$. For

a temperature $t$, $t/t_s$ gives a measure of the relative entropic contribution to the stiffness, as compared to the enthalpic contribution.

In this section, the sequence-dependence is studied at the base-pair level : we include the possible influence of the neighbor bp by separating the different *trinucleotides*, as displayed in Table 1.3.

| Trinucleotide | Occurrences |
|---|---|
| A**A**A | 2 |
| A**A**C | 2 |
| C**A**A | 3 |
| G**A**T | 2 |
| T**A**G | 3 |
| T**A**T | 2 |
| A**C**A | 2 |
| C**C**C | 9 |
| G**C**G | 9 |
| T**C**T | 2 |

TABLE 1.3 – 10 available trinucleotides sequences : 6 different environments of the A-T bp and 4 different environments of C-G

We recall that the *covariance* coefficients are plotted on gray background, so that they can be easily distinguished from the stiffness elements, and that they are normalized by the temperature.

### 1.4.1 Data analysis

**Sequence-averaged analysis** The translational *stretch* parameter exhibits limited sequence-dependence, and can be analyzed with some success on the whole dataset. In this case, we computed the error estimates from the deviation of the values obtained at the different bp positions.

The upper panel of Fig. 1.10A shows the evolution of the stretch-stretch stiffness as computed on the sequence-averaged dataset. The value of $t_s = 2.21 \pm 0.35$ indicates that the entropic contribution is considerable : at room temperature, the stiffness is already divided by 2 wrt the enthalpic stiffness.

For the other degrees of freedom however, the distribution varies too much with respect to the sequence for a similar analysis : the approximation of a multimodal distribution by a unique covariance matrix is here artificial. The signature of this artificial construction is that the features are dominated by apparent noise, which is due to structural heterogeneity rather than thermal fluctuations, as shown on Fig. 1.10A (lower panel) for the opening-opening diagonal element.

Next, we look at the sequence. We can divide the data into two groups, whether the considered bp is A-T or C-G : two parameters (shear and opening) exhibit relatively homogeneous values within these two groups. The influence of the neighboring sequence will be taken into account afterwards, which will further divide the dataset into the 10 trinucleotides listed in Table 1.3.

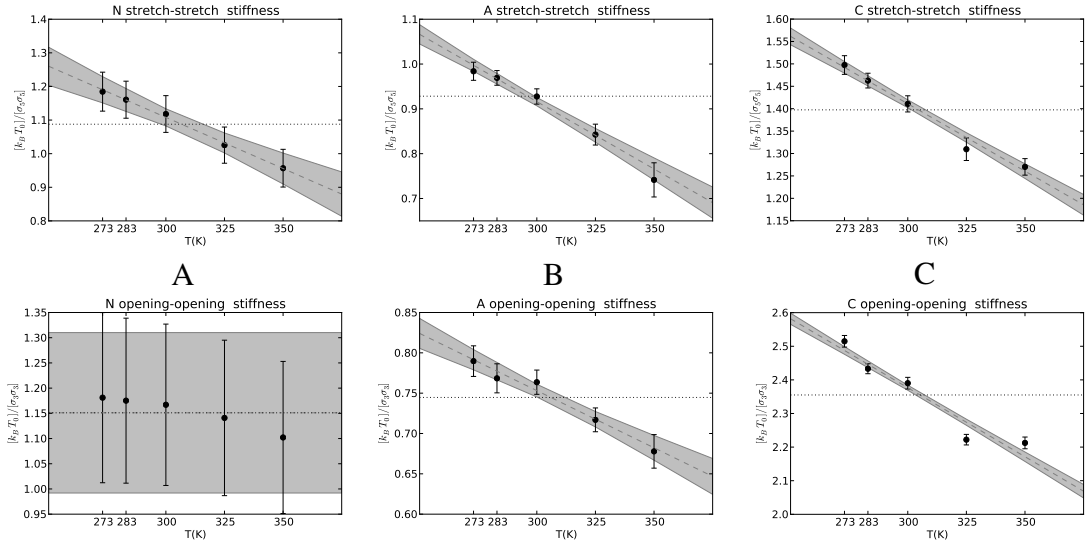


FIGURE 1.10 – Regression of the stiffness values for the stretch-stretch and opening-opening diagonal term : (A) on the sequence-averaged dataset ; (B) on the A-T nucleotides ; (C) on the C-G nucleotides. The stretch parameter exhibits a sequence-independent temperature evolution, in contrast to the opening flexibility. However, the separate analysis for the two groups of bp show that it is just a question of heterogeneity. For the former, the entropic contribution to the stiffness is similar for both sequences, while the enthalpic part is different, thus a different $t_s$. The opening parameter, on the other hand, is around three times stiffer for the C-G bp at room temperature, but its entropic contribution is also three times larger, and therefore the two bases become unstable at around the same temperature in this direction.

**A-T and C-G base-pairs** Interestingly, the *stretch-stretch* diagonal term which had been resolved in the sequence-neutral analysis, is refined into two well-separated behaviors depending on the considered bp. While the stiffness constant at $T_0$ of the C-G bp is $\sim 50\%$ larger than that of A-T, the entropic stiffness constant is similar $k_s = 66 \pm 9 [k_B T_0]/\mathring{A}^2$, hence two different values $t_s^{C-G} = 2.57 \pm 0.16$ and $t_s^{A-T} = 2.02 \pm 0.15$ (see Fig. 1.10).

A yet different behavior is that of the *opening* parameter, which exhibits considerable temperature-dependence once the A-T and C-G bp are separated. At room temperature, the stiffness of the C-G bp is three times stronger than that of the A-T, which can be attributed to the different number of hydrogen bonds. But here, the entropic contribution is also $\sim 3$ times higher for G-C, so that its *relative* weight is similar, and thus the two bp are predicted to become unstable at the same temperature $t_s \simeq 2.8 \pm 0.2$ in that direction.

Finally, *shear* exhibits limited (but nonzero) sequence-dependence for a given nucleotide. Both $k_g^0$ and $k_s$ are rather different between the sequences, with values $t_s^{C-G} = 4.0 \pm 0.4$ and $t_s^{A-T} = 6.8 \pm 2.6$. For the two latter parameters, the relatively large error bars reflect the remaining heterogeneity among different sequences, *i.e.* the influence of the flanking nucleotides that we analyze now.

**Sequence-specific analysis** In the previous paragraphs, we noticed that some degrees of freedom exhibit a sensitivity to temperature. What about the others ? Is there a sequence-dependent refinement for the entropic contributions we have computed so far ? To answer these questions, let us analyze the dataset for each dinucleotide separately.

The detailed error analysis presented in Section 1.3 shows that there are at least two times-

cales present in the data : a rapid motion of characteristic time $\tau \simeq 50-100$ps that dominates the variance, and an equilibration time $\tau \simeq 1-10ns$. The main contribution to the error comes from the rapid motion, which can be efficiently resolved in our data. However, because our sampling is limited wrt the slower motion, it is sometimes difficult to estimate this error with a better accuracy than $\sim 30\%$. In some cases, a very slow equilibration time was detected, in which case the error estimation is more delicate, and we increased our estimations for security.

We determine which parameters present a detectable entropic contribution to the stiffness by applying a criterion, which compares the accuracy of the enthalpic+entropic model with that of the pure enthalpic one. Again, the details are given in the Analysis section 1.3. The results obtained at this level confirm and go beyond those of the previous step : the test is positive for all nucleotides with respect to the opening, shear and stretch diagonal elements respectively. In most cases, the separate analysis of the different sequences provides more precise information. For instance, for the opening angle, the obtained $t_s$ are all between 2 and 3, and the error bars are sufficiently small to distinguish the values of the different sequences. Tables 1.4 and 1.5 give the detailed results for opening and stretch respectively. More detailed values can be found in Appendix, Section 1.7.5.

| Seq | $k_g^0$ | $k_s$ | $t_s$ |
|---|---|---|---|
| AAA | $\mathbf{1.088 \pm 0.003}$ | $\mathbf{0.826 \pm 0.028}$ | $\mathbf{2.318 \pm 0.043}$ |
| AAC | $\mathbf{1.083 \pm 0.002}$ | $\mathbf{0.777 \pm 0.021}$ | $\mathbf{2.394 \pm 0.037}$ |
| CAA | $\mathbf{1.001 \pm 0.003}$ | $\mathbf{0.828 \pm 0.027}$ | $\mathbf{2.209 \pm 0.037}$ |
| GAT | $\mathbf{0.984 \pm 0.004}$ | $\mathbf{1.038 \pm 0.054}$ | $\mathbf{1.948 \pm 0.049}$ |
| TAG | $\mathbf{0.929 \pm 0.003}$ | $\mathbf{0.976 \pm 0.037}$ | $\mathbf{1.951 \pm 0.036}$ |
| TAT | $\mathbf{0.931 \pm 0.004}$ | $\mathbf{0.946 \pm 0.039}$ | $\mathbf{1.984 \pm 0.04}$ |
| ACA | $\mathbf{1.478 \pm 0.004}$ | $\mathbf{0.977 \pm 0.04}$ | $\mathbf{2.513 \pm 0.06}$ |
| CCC | $\mathbf{1.656 \pm 0.002}$ | $\mathbf{1.097 \pm 0.029}$ | $\mathbf{2.51 \pm 0.04}$ |
| GCG | $\mathbf{1.457 \pm 0.002}$ | $\mathbf{1.03 \pm 0.02}$ | $\mathbf{2.415 \pm 0.027}$ |
| TCT | $\mathbf{1.622 \pm 0.003}$ | $\mathbf{1.08 \pm 0.042}$ | $\mathbf{2.502 \pm 0.058}$ |

TABLE 1.4 – Fitted parameters for (stretch-stretch)

| Seq | $k_g^0$ | $k_s$ | $t_s$ |
|---|---|---|---|
| AAA | $\mathbf{0.906 \pm 0.004}$ | $\mathbf{0.498 \pm 0.034}$ | $\mathbf{2.821 \pm 0.119}$ |
| AAC | $\mathbf{0.917 \pm 0.003}$ | $\mathbf{0.448 \pm 0.027}$ | $\mathbf{3.047 \pm 0.121}$ |
| CAA | $\mathbf{0.847 \pm 0.004}$ | $\mathbf{0.477 \pm 0.035}$ | $\mathbf{2.775 \pm 0.124}$ |
| GAT | $\mathbf{0.817 \pm 0.005}$ | $\mathbf{0.612 \pm 0.069}$ | $\mathbf{2.335 \pm 0.147}$ |
| TAG | $\mathbf{0.763 \pm 0.004}$ | $\mathbf{0.417 \pm 0.03}$ | $\mathbf{2.831 \pm 0.127}$ |
| TAT | $\mathbf{0.807 \pm 0.004}$ | $\mathbf{0.358 \pm 0.034}$ | $\mathbf{3.252 \pm 0.204}$ |
| ACA | $\mathbf{2.857 \pm 0.008}$ | $\mathbf{1.467 \pm 0.072}$ | $\mathbf{2.948 \pm 0.093}$ |
| CCC | $\mathbf{2.843 \pm 0.003}$ | $\mathbf{1.178 \pm 0.03}$ | $\mathbf{3.414 \pm 0.06}$ |
| GCG | $\mathbf{2.412 \pm 0.004}$ | $\mathbf{1.347 \pm 0.035}$ | $\mathbf{2.791 \pm 0.045}$ |
| TCT | $\mathbf{2.717 \pm 0.007}$ | $\mathbf{1.141 \pm 0.07}$ | $\mathbf{3.38 \pm 0.143}$ |

TABLE 1.5 – Fitted parameters for (opening-opening)

In contrast, the two other angular parameters exhibit very little or no detectable entropic contribution to the stiffness (1 and 3 positive results). The last translational parameter, stagger,

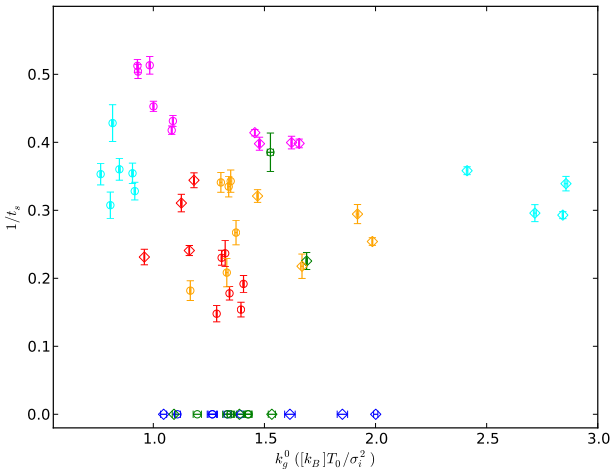


FIGURE 1.11 – Results for the intra-bp diagonal parameters : relative entropic contribution (as measured by $1/t_s$) against relative stiffness at room temperature $k_g^0$. Here $k_g^0$ is expressed in units of the sequence-averaged stiffness of the considered degree of freedom, at $T_0 = 300K$ : the values of the different parameters should not be compared (see Table 1.2 for the average standard deviations $\sigma_i$ for the different degrees of freedom). The different colors correspond to the different parameters : buckle (blue), propeller (green), opening (cyan), shear (red), stretch (magenta), stagger (orange). C-G nucleotides are plotted with a circle, A-T with diamonds. The parameter with largest relative entropic weight is stretch, then opening, *i.e.* the stiffest parameters. For opening (cyan), the two types of bp have very different stiffnesses, but the entropic contributions are similarly related, which gives comparable spinodal temperatures. For stretch (magenta), the stiffness is also sequence-dependent, but the enthalpic part is much more regular.

is also strongly temperature-dependent for all sequences. The results are presented on Fig. 1.11, where the different diagonal parameters are plotted together to illustrate the different behaviors.

**Minimal set of parameters** We have tested if the number of parameters can be reduced by considering a common value of $t_s$ for the different sequences, while allowing different $k_g^0$. In each case, we use again a criterion to compare the predictive power of the reduced parameter set, as compared to the complete sequence-dependent parametrization, as described in detail in Sec. 1.3.3. We find that the comparison is always in favor of the detailed parameter set, even for the cases like the stretch-stretch or opening-opening diagonal elements, where the estimated $t_s$ for the different sequences are rather close, while the value at $T_0$ can be rather different. This is reflected in the estimated error bars on the spinodal temperature $t_s$, which are sufficiently small to separate the different trinucleotides.

Up to now we have discussed only the fits of the stiffness values. We checked that the model reproduces the behavior of the covariance with temperature, by inverting the stiffness matrices under the assumption that the errors of the different elements are not correlated. This sometimes results in overestimating the errors in the covariance, which may compensate the possible underestimation of those for the stiffness.

The estimated values agree satisfactorily with the measured points for the diagonal terms, as is illustrated in Fig. 1.12 for the CAA trinucleotide. For buckle and propeller, no contribution of entropy can be detected in the data, and the fitted model (shaded area) is close to a purely enthalpic stiffness. The small remaining slope in the covariance plot is the result of the degrees of freedom being mixed in the matrix inversion operation. In this example, the procedure found

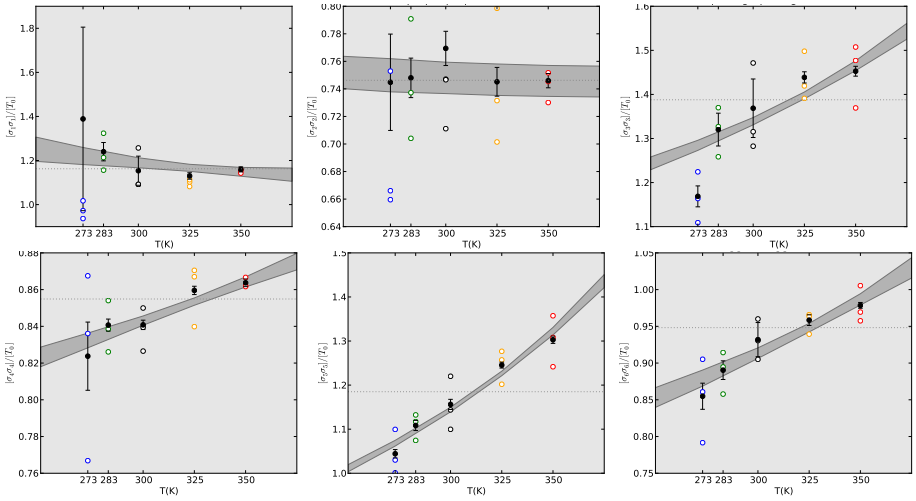

FIGURE 1.12 – Diagonal elements of the covariance matrix normalized by temperature, for the trinucleotide CAA : buckle, propeller twist and opening (upper row), shear, stretch and stagger (lower row). The individual values (colored dots) correspond to the different positions of the trinucleotide in the oligomers. The error bars are computed by the block averaging method (see previous section), while the mean values (black dots) are computed on the complete dataset. The shaded are is the predicted covariances, obtained from the inversion of the fitted stiffness matrix, where the errors are assumed to be independent. The purely enthalpic model is shown as a dotted line.

a very large correlation time for buckle at 273K, as shown by the large error bar. For opening and for the three translational parameters, the substantial entropic contribution to the covariance is well reproduced.

### 1.4.2 Interpretation : the path to the melting transition and the spinodal decomposition

**Path to the melting transition** Our results demonstrate the presence of an important entropic contribution in the flexibility of the base-pair (bp), and show that this contribution affects some degrees of freedom (stretch, opening) more that others. This contribution is very likely to have consequences for the stability of the double-helical phase. The melting transition itself is beyond the scope of our study, because we have no information on the properties of the ss phase and the thermodynamics of bubble formation. And yet, at least as a qualitative observation, the degrees of freedom most sensitive to temperature are likely to be also involved in the phase transition. This impression is consistent with the striking observation that base-pair stretching is the parameter where the entropic effect is strongest and most generic. We estimate that the A-T and G-C base-pairs have a similar entropic contribution, but differ by the enthalpic stiffness constant, which seems also qualitatively compatible with the well-known superior thermodynamic stability of the G-C bp. If the entropic contribution to the stiffness is indeed indicative of the *initial direction of the melting path*, then we estimate that simultaneously to stretching, the bases get mostly distorted in the opening direction, *i.e.* in the base-pair plane, and very little in the other angles. This is probably true only as long as the base-pairs remain stacked. It also raises the question, whether the internal base-pair or the stacking free energies are dominant for melting. This question will be discussed after looking at the step parameters in the next section.

To illustrate that the melting transition cannot be simply related to our data, we have tested,

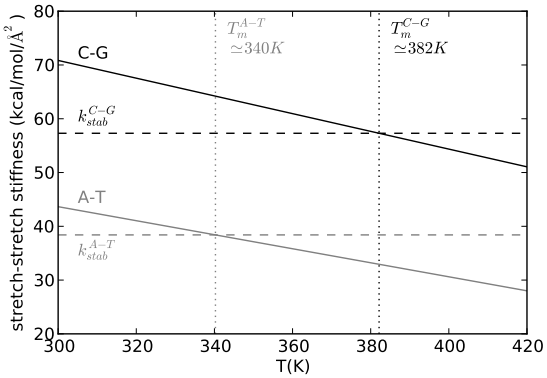


FIGURE 1.13 – Comparison of our temperature-dependent stretch-stretch elasticity with the estimated melting temperatures associated with the hydrogen-bonding contribution [SantaLucia 1998], in a simplistic model where melting occurs when the stiffness drops below a limiting stability value. The experimentally-derived temperatures cannot be reproduced by a single value for the two bp types.

whether the stability of the bp could be simply described in terms of a limit stiffness value for the stretch direction, assuming that if the bp becomes too floppy, the dh may locally melt. We compared our temperature-dependent values, averaged over the base-pair type A-T/C-G, with the estimated melting temperatures associated to the hydrogen bonding contribution only [SantaLucia 1998]. From the values given in this paper, this approach could be easily extended to include the stacking interaction and comparing the values for all bps. Fig. 1.13 shows, that the experimentally-derived temperatures with our salt concentration cannot be simply reproduced using a common value for the limit stiffness : the C-G bp requires a much higher value. One could consider to test other equally simple constructions, for instance where there is a maximum distance between the bases, but we didn't investigate it by lack of time.

**Spinodal temperature** The elastic model is the first order approximation of the energy landscape : from this approximation, one can compute the extrapolation to the spinodal decomposition, which happens when the considered phase becomes unstable (see section 1.1). This is to be contrasted with the melting temperature, which characterizes the equilibrium between two phases : for $T_m < T < T_s$, the condensed phase (here the dh) can still exist as a metastable state, and not for $T > T_s$. As an example, the spinodal temperature of liquid water at atmospheric pressure is estimated at $T \simeq 300$°C [Eberhart and II 1985].

The spinodal decomposition is estimated at the temperature where the total stiffness matrix has a null eigenvalue : the fluctuations in the direction of the associated eigenvector become considerable, *i.e.* this vector is the predicted direction of decomposition. The spinodal temperature was computed by a numerical procedure, where the stiffness matrix is diagonalized at different temperatures to find the interval where the first eigenvalue becomes negative, and the operation is repeated iteratively on this interval to increase the accuracy. The computed values are displayed in the first column of Table 1.6. They are between 1.7 and 2.2 for all sequences. As noticed before, it is an interesting feature that A-T and G-C are predicted to become unstable at similar temperatures, although they have very different stiffness values for the most important parameters. The melting vectors are displayed in the table : not surprisingly, the bp are destabilized in the stretch-stretch direction, with a simultaneous partial opening.

| | $t_s$ | buckle | propel | opening | shear | stretch | stagger |
|---|---|---|---|---|---|---|---|
| AAA | 2.034 | -0.287 | 0.337 | -0.56 | 0.12 | -0.606 | -0.33 |
| AAC | 2.179 | -0.019 | 0.149 | -0.05 | -0.014 | -0.856 | -0.492 |
| CAA | 2.102 | 0.067 | -0.051 | 0.456 | -0.057 | 0.873 | 0.139 |
| GAT | 1.743 | 0.109 | 0.156 | -0.427 | 0.04 | 0.799 | 0.376 |
| TAG | 1.848 | 0.129 | -0.072 | -0.065 | -0.032 | -0.955 | -0.246 |
| TAT | 1.802 | 0.075 | 0.094 | -0.479 | -0.072 | -0.84 | -0.213 |
| ACA | 1.753 | 0.096 | 0.082 | -0.565 | -0.36 | -0.707 | -0.188 |
| CCC | 1.877 | -0.04 | -0.14 | 0.554 | -0.23 | 0.782 | 0.083 |
| GCG | 2.034 | 0.002 | 0.213 | -0.521 | 0.039 | -0.792 | -0.235 |
| TCT | 1.802 | -0.024 | 0.141 | -0.569 | 0.361 | -0.711 | -0.143 |

TABLE 1.6 – Spinodal decomposition of the base-pairs : temperature and direction of instability. The components of the eigenvector are expressed in reduced units of the different parameters, where they have comparable values at $T_0$. It therefore represents their respective weight in the direction of instability.

### 1.4.3 Mean values

So far, we have considered only the stiffness/covariance, *i.e.* the second moment of the distribution, which is modeled according to the first part of Eq. 1.16. We now look at the evolution of the equilibrium values, by using a very similar approach (see previous section).

The parameters where temperature induces an average displacement are also those where the stiffness changes. The average *stretch* of all base-pairs increases with temperature. Not surprisingly, this effect is generally stronger for the A-T bp, which have already a larger stretch at room temperature. The mean *opening* also increases with temperature, but mainly for the G-C bp. The other angles do not exhibit a systematic effect. The detailed values can be found in Appendix.

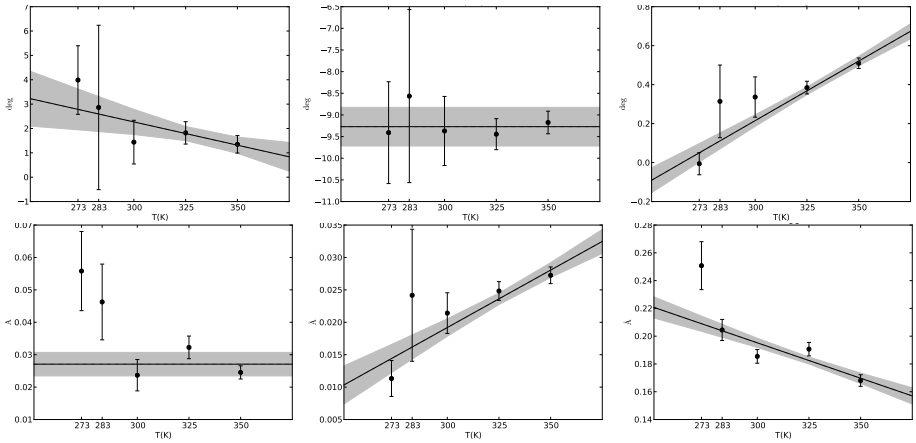


FIGURE 1.14 – Fits of the internal equilibrium orientations and positions for the GCG trinucleotide : buckle, propeller twist, opening (upper row), shear, stretch, stagger (lower row)

## 1.5 Results II : base-pair step elasticity

In this section, we model the data as has been done for the intra-bp parameters. We show that entropy contributes to the elasticity, especially for tilt, roll and rise. *i.e.* the parameters which dominate the large-scale bending flexibility. This effect is weaker than in the previous case, but detectable, and our analysis allows to parametrize a T-dependent rigid base-pair (rbp) model, following the model described in Section 1.1. From these temperature-dependent parameters, we can use coarse-graining relations to estimate the entropic contribution in the dh persistence length.

### 1.5.1 Parametrization of a T-dependent rigid base-pair model

**Fitting the stiffnesses** For the step parameters, we divide the data according to the 11 tetranucleotide sequences showed in Table 1.1, p. 43.

The values of the stiffnesses are less uniform than for the intra-parameters, and so is the effect of temperature. The largest contribution is found for the two bending angles, tilt and roll, and for the translational parameter rise (and to a lesser extent, slide). More surprisingly, the twist and shift stiffnesses exhibit a very low sensitivity to temperature.

The values of $k_g^0$ and $t_s$ for the *roll-roll* stiffness term and for the different sequences are given in Table 1.7. As one can see, the values are less regular than for the intra-parameters, and the noise is more important, but some entropic contribution is detectable for most dinucleotides, yielding values $2 < t_s < 4$. No immediate correlation can be found between the estimated entropic content and the sequence type, for instance between purines and pyrimidine steps.

| Seq | $k_g^0$ | $K_s$ | $t_s$ |
|---|---|---|---|
| **AA** | $\mathbf{1.905 \pm 0.017}$ | $\mathbf{0.712 \pm 0.16}$ | $\mathbf{3.677 \pm 0.586}$ |
| **AA1** | $\mathbf{1.953 \pm 0.013}$ | $\mathbf{0.547 \pm 0.099}$ | $\mathbf{4.571 \pm 0.632}$ |
| **AC** | $\mathbf{2.205 \pm 0.018}$ | $\mathbf{1.296 \pm 0.157}$ | $\mathbf{2.701 \pm 0.197}$ |
| AG | $1.82 \pm 0.023$ | $0.0 \pm 0.0$ | $\infty$ |
| **AT** | $\mathbf{2.218 \pm 0.019}$ | $\mathbf{0.919 \pm 0.179}$ | $\mathbf{3.414 \pm 0.459}$ |
| **CA** | $\mathbf{1.766 \pm 0.023}$ | $\mathbf{0.336 \pm 0.185}$ | $\mathbf{6.26 \pm 2.843}$ |
| **CG** | $\mathbf{1.892 \pm 0.015}$ | $\mathbf{1.7 \pm 0.118}$ | $\mathbf{2.113 \pm 0.071}$ |
| **GA** | $\mathbf{1.784 \pm 0.014}$ | $\mathbf{0.38 \pm 0.118}$ | $\mathbf{5.7 \pm 1.433}$ |
| **GC** | $\mathbf{2.095 \pm 0.008}$ | $\mathbf{0.917 \pm 0.065}$ | $\mathbf{3.285 \pm 0.158}$ |
| **GG** | $\mathbf{1.95 \pm 0.004}$ | $\mathbf{1.026 \pm 0.042}$ | $\mathbf{2.9 \pm 0.076}$ |
| TA | $1.508 \pm 0.024$ | $0.0 \pm 0.0$ | $\infty$ |

TABLE 1.7 – Estimated parameters for the (roll-roll) temperature-dependent stiffness. The lines with a thin typeface are those where no entropic contribution was detected.

The detailed results, with the fits of the diagonal stiffness elements for all sequences, are shown in Appendix, Section 1.7.6.

**Covariances** Fig. 1.15 shows the datapoints and the model prediction for the covariance diagonal elements of the GG dinucleotide, the one with most data-points. We do observe an evolution with temperature for all degrees of freedom, which is satisfactorily reproduced by the

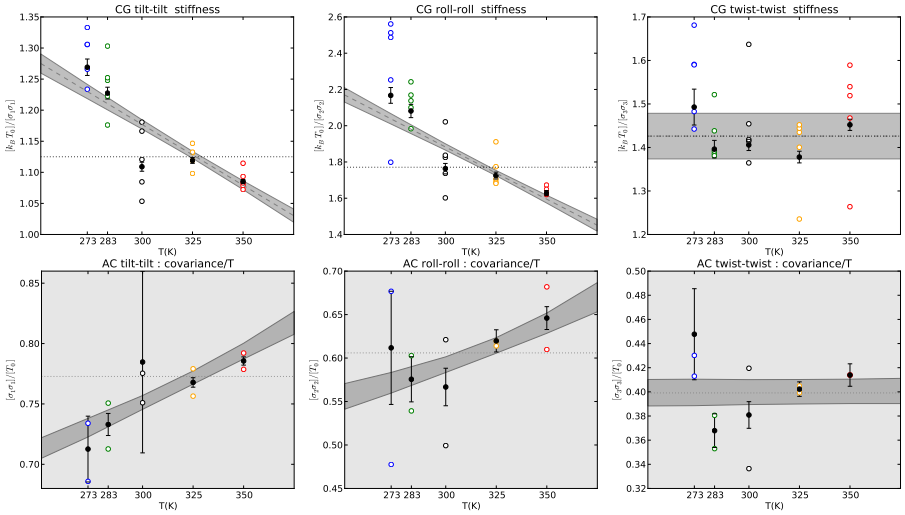


FIGURE 1.15 – Regression of the data for the three angular diagonal terms of the CG dinucleotide : (left) tilt-tilt ; (center) roll-roll ; (right) twist-twist. The colored dots correspond to the values computed at the different positions of the CG step in the oligomer. For some datapoints, a large correlation time was detected, and the error bar is therefore very large.

linear model of the stiffness. Here in most cases the effect is close to the level of statistical noise, except for the roll parameter.

The conclusion we can draw is that the entropic effect on base-pair step elasticity is more limited than inside the base-pair. It may also be affected by slower equilibration times, which would require a greater computational effort to resolve. However, it must be noted that the three parameters most influential in terms of large-scale bending are also the ones where the effect of temperature is strongest, and in the next section we will compute estimates of the temperature evolution of the persistence length. Before that, we investigate the evolution of the equilibrium values.

**Equilibrium values** It is a question of considerable interest, whether the base-pair stacking of DNA changes spontaneously with temperature. An example of the evolution of the step parameter equilibrium values is shown in Fig. 1.16, for the GG dinucleotide.

The only parameter where a clear and sequence-independent tendency emerges out of statistical noise is rise, which typically increases of $2-3\%$ in the considered temperature interval.

In particular, the equilibrium twist angle does not exhibit any systematic change with temperature, decreasing by 2-4° for some sequences (e.g. AA, GC, GA) while remaining constant for others (GG). This probably indicates the absence of a detectable spontaneous change in supercoiling for organisms living at high temperature. The latter may therefore be related to the presence of proteins in charge of this specific task (see the introduction of this chapter) or to more subtle effects involving cations.

The detailed plots for all parameters and sequences are given in Annex, Fig. 1.26, p. 85.

Finally, as a result of the analysis, we have parametrized a T-dependent rbp model of DNA, where the entropic contributions are sequence-dependent. For the AA dinucleotide where we had two different contexts, we arbitrarily chose the values computed on the “AA1” dataset, where the mean values are closer to those reported in the crystal structures, and where we had more datapoints. Altogether, the model includes 156 nonzero “entropic” parameters, in addition

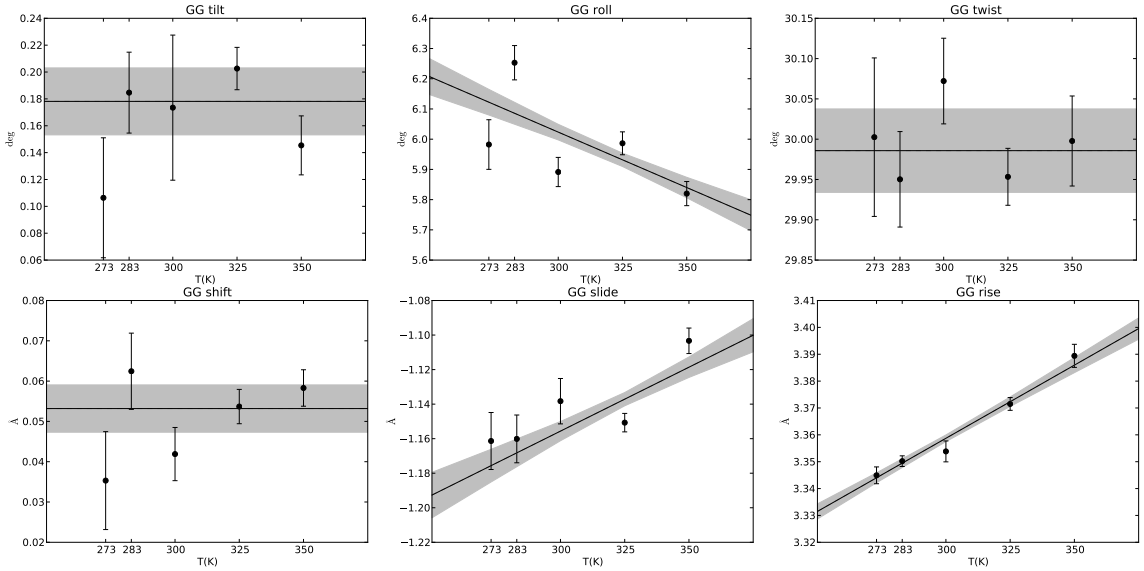


FIGURE 1.16 – Fits of the equilibrium orientations and positions for the GG dinucleotide (step parameters)

of the 270 parameters required for describing the elasticity at a single temperature.

### 1.5.2 Flexibility of the coarse-grained WLC model

**Introduction** Measurements of the bending persistence length of DNA exhibited a stronger temperature-dependence than expected from a purely enthalpic model of elasticity. In [Theodorakopoulos and Peyrard 2012], the authors assume that the stiffness of the double helix is constant, and show that the observed decrease can be partly reproduced by the spontaneous bubbles occurring in the premelting stage : see Fig. 1.1. In the previous paragraphs, we showed that in the double-helical phase, the stiffness at the bp and bps level is not independent of temperature, but decreases due to a detectable entropic contribution.

In the MD trajectories, the DNA remains in the double strand state, and the effect of bubbles is not included in these values. It is unclear if this is for purely kinetic reasons, or a consequence of the force fields. Assuming that the sampled energy landscape is representative of that of the real molecule in the double-strand phase, we use coarse-graining relations to estimate the values of the persistence length, from the model computed at the bps level.

Here, we neglect the influence of the intra-bp deformations on the large-scale elasticity, and consider only the step parameters. The most naive way of computing the persistence length is an immediate extrapolation from the mean variance of the tilt and roll angles (see Section 0.2.3) : the bending persistence length is then equal to :

$$l_p = \frac{2b}{\langle(\tau-\tau_0)^2\rangle + \langle(\rho-\rho_0)^2\rangle} \tag{1.23}$$

where $\tau$ and $\rho$ are the tilt and roll angles respectively, $\tau_0$ and $\rho_0$ are their mean value, and $b \simeq 0.34$nm is the average rise of the considered dinucleotide.

This approximation would be valid only in the case where all bps were parallel and aligned on the helical axis. The real molecule deviates from such an ideal B-DNA helix, and all step

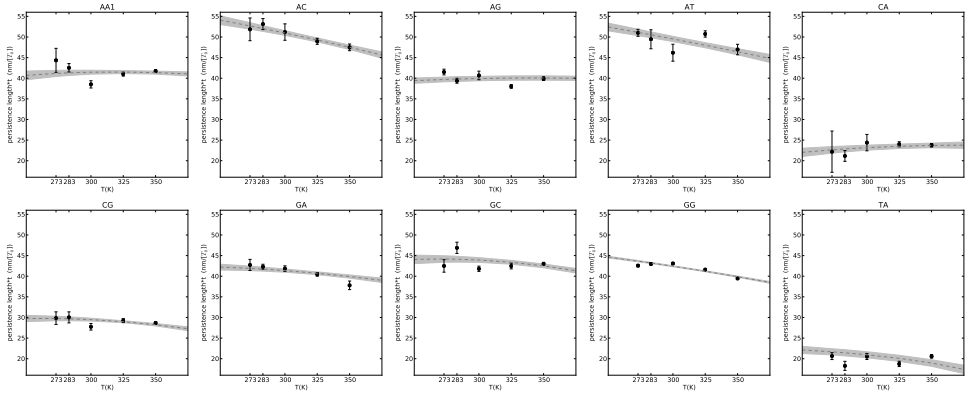


FIGURE 1.17 – Pseudo-persistence lengths obtained from the sequence-dependent elastic parameters extracted from the simulations, multiplied by the temperature. In this representation, a purely enthalpic stiffness gives a constant value, whereas an entropic contribution generates a decrease with temperature. The temperature dependence is visible for some dinucleotides, but remains much smaller than the sequence-dependent dispersion. The values of the persistence length are computed rigorously from the complete step parameters (see text), while the error bars are estimated from the main contributions (mean rise, roll and tilt angles) in an approximate calculation.

degrees of freedom contribute to the overall bending. We therefore use a coarse-graining calculation [Becker and Everaers 2007], where the internal bp deformations are neglected, but the deviations from the perfect B-DNA conformation are taken into account. To estimate the error bars on the computed values, we used the naive version of the calculation, Eq. 1.23. Note that because in our case the deformations from the ideal helix remain small, the deviations between both methods are typically less than $\sim 5\%$.

From the sequence-dependent enthalpic and entropic stiffness matrices obtained by fitting the simulation points, the computation provides sequence-dependent "pseudo-persistence lengths". Here the prefix "pseudo" refers to the fact that for most sequences, sequence continuity forbids the existence of oligomers composed of these dinucleotides only ; one must understand the figures as illustrating the respective contributions of the different sequences to the sequence-averaged persistence length. Note that, as in Fig. 1.1, we plot the persistence length multiplied by temperature, which directly shows the entropic contribution to the large-scale bending stiffness, as explained previously.

As noted at the bps-level, the temperature dependence is detectable only for some sequences, and the variations of some sequences are very noisy, in which case the fitted model sometimes exhibits a small trend, which was absent in the data. This effect comes from the inversion of the stiffness matrix, as noticed during the fitting procedure, and it is negligible for the sequence-neutral values that we compute now.

The sequence-neutral persistence length is computed by averaging the mean rise and the angle fluctuations (covariance matrix) over the 16 possible dinucleotides. The resulting value of the persistence length is shown on Fig. 1.18(left) together with the datapoints from [Theodorakopoulos and Peyrard 2012]. As can be seen, the computed values are significantly lower (around 10nm) than the experimental ones. This effect is not a consequence of irregularities in our simulations, since the same effect was already observed in [Becker and Everaers 2007] for different parameter sets. We compared the computed values with a hybrid parameter set

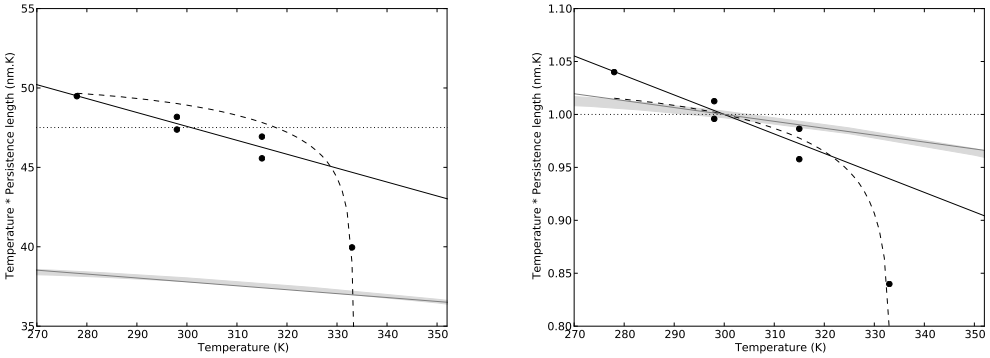


FIGURE 1.18 – Sequence-averaged persistence length, multiplied by temperature. Dots : experimental points. Shaded area : MD-derived values of the persistence length, within error bars. The dark gray line is the best linear fit of the MD-derived values. Dashed line : predicted bubble contribution [Theodorakopoulos and Peyrard 2012]. Solid line : best fit of the experimental values, where the last point is ignored. Dotted line : best fit, enthalpic model : **(left)** The values computed from the simulations are significantly lower than the experimental values from [Geggier et al. 2011], as already observed in [Becker and Everaers 2007]. **(right)** To compare the relative contribution of entropy, the values at all temperatures are rescaled by that at 300K. For temperatures in the range 273-310K, the entropic contribution in the dh is about the same as the contribution from bubbles : both are close to half the experimental value, which immediately suggest the construction of a hybrid model, where the contributions are likely to be additive

where the mean conformations at 300K are taken from the crystallographic data : the results are consistent. Values from the naive computation are equally underestimated. It is unclear, whether this is a general feature of the MD fluctuations or a consequence of the calculation method : we assume that this issue does not influence the *temperature dependence* of the predicted values, and we therefore rescale them by their value at 300K (right).

At the sequence-neutral level, the temperature dependence is limited but emerges out of the estimated errors. Interestingly, the best linear fit of the data suggests that for temperatures 273K<T<310K, the two entropic contributions, in the dh and by bubble formation, are of the same magnitude. Qualitatively, they are both close to half the experimental slope, which immediately suggests the construction of a hybrid model, where the temperature dependence of the dh stiffness is included together with the possibility of bubble formation. In this case, we expect the two contributions to sum up, and result in a faster decrease in this temperature range. For the larger temperatures, not surprisingly, the elastic contribution remains approximately linear, while the bubble contribution explodes with the fraction of melted DNA.

## 1.6 Conclusion and outlook

In this chapter, we have studied the fluctuations of the double helix over a wide range of temperatures. We have established that the elasticity of dsDNA contains a significant entropic contribution : as a result, the temperature dependence of some degrees of freedom is substantially stronger than generally assumed from a purely enthalpic model.

**Base-pair fluctuations and melting transition** We found that this entropic part of the stiffness is more important for the internal base-pair fluctuations than for the step parameters, espe-

cially for the base-pair stretching and opening degrees of freedom, and in a sequence-dependent way. These observations give the qualitative view that in the premelting stage, the base-pairs exhibit substantial fluctuations in these directions before breaking. The path toward the melting transition is therefore probably initially controlled by the internal base-pair deformations, which increase until the unstacked state becomes more favorable. In turn, the transiently broken and unstacked bp influences the free energy landscape of the neighbor bp, favoring a collective opening of the double helix. Recent experiments demonstrated that this cooperative effect can extend 10bp away and possibly more [Cuesta-López et al. 2011].

**Molecular mechanisms for entropy-driven elasticity** Coming back to the fluctuations of the dh, our study does not provide the molecular mechanism behind this entropic contribution to elasticity. In [Travers et al. 2012], the authors suggest that the apparent elastic properties of dsDNA could be modified by the water molecules penetrating into the grooves, where they form a ladder structure, constraining the molecule in a bent conformation, and reducing the transverse motions of the bases. This structuration could be destroyed at higher temperature, with strong consequences on the flexibility of the bases, and possibly in a sequence-specific way. As a future extension of this aspect of the study, we may test this hypothesis explicitly by computing the water density in the helical frame, by using an adapted extension of the Curves software specifically designed for such calculations [R. Lavery, private communication]. The same analysis could then be carried for the ion distributions.

**Base-pair step parameters and nucleosome association free energies** The second results concern the temperature-dependence of the base-pair step elastic parameters. Because of the limited temperature dependence observed in this case, we spent some effort in estimating the range of validity of the computed results. The computed error bars are the typical uncertainty in the values of the last generation of MD-derived parameters, which are used in coarse-grained models like the rigid base-pair model. This model can be used to estimate the sequence-dependent nucleosome free energies, where the most common method is to thread sequences on a rigid structure derived from the crystallographic models. This computation is not very satisfactory, as discussed in the next chapter, but the sequence-dependent affinities are at least qualitatively reproduced [Deniz et al. 2011]. With our implementation of the rigid base-pair model of DNA, we can extend this study by systematically comparing the binding affinities (i) for different sequences at the tetranucleotide level, by including the complete "ABC" parametrization ; (ii) at different temperatures, with our entropic stiffness matrix ; and (iii) estimate the significance of the estimated sequence- and temperature-dependent variations by including the uncertainties of the parameter set. The latter step involves the MC generation of a range of parameter sets following the method discussed in this study, and computing the estimated nucleosome free energy for each of them.

## 1.7 Appendix

### 1.7.1 Direct estimation of thermodynamic quantities

#### (a) Phase transition ?

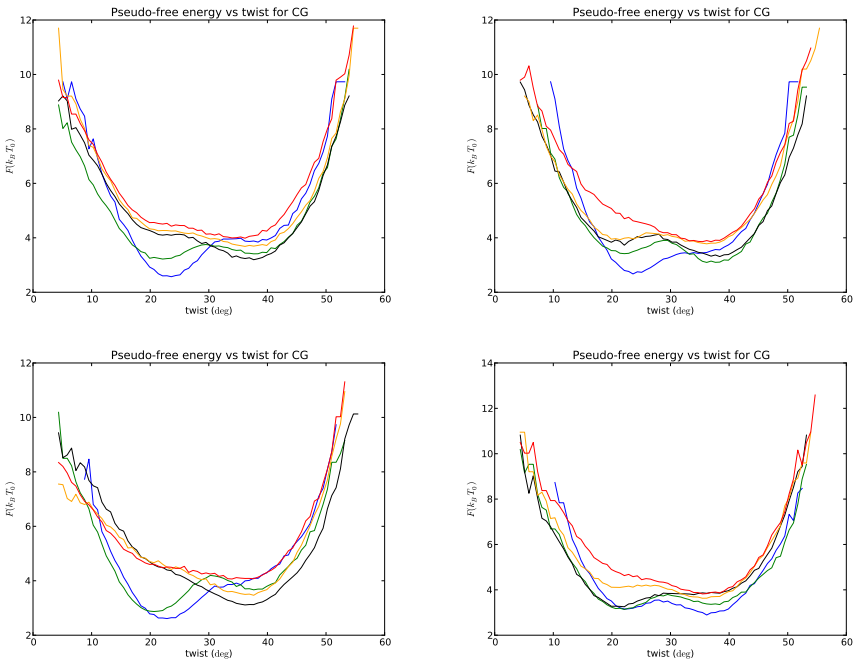


FIGURE 1.19 – The twist-projected free enthalpy computed from the distributions at four positions of the dinucleotide CG show that the "phase transition" appearance of the whole data may be an artifact of insufficiently equilibration at the lowest temperature.

#### (b) Entropy and enthalpy : direct calculation

In principle, from the estimated free enthalpies estimated in Section (b), one can compute the entropy and the enthalpy :

$$\begin{cases} S(q,T) = \frac{\partial G}{\partial T}(q,T) \\ H(q,T) = G(q,T) + TS(q,T) \end{cases} \tag{1.24}$$

For a discrete set of temperatures, the partial derivatives are simply replaced by differences :

$$\begin{cases} S(q,T_s) = \frac{G(q,T_1) - G(q,T_2)}{T_2 - T_1} \\ H(q,T_s) = \frac{G(q,T_1) + G(q,T_2)}{2} + T'S(q,T') \end{cases} \tag{1.25}$$

where $T' = (T_1 + T_2)/2$. However, the poor sampling prevents any direct calculation. Again, one may compute an approximate value from the pseudo-free enthalpy, as shown in Fig. 1.20.

Even here, the differentiation effectively amplifies the noise, and the values cannot be quantitatively compared. We circumvent this problem in the analysis by considering the Gaussian approximation of the distribution.

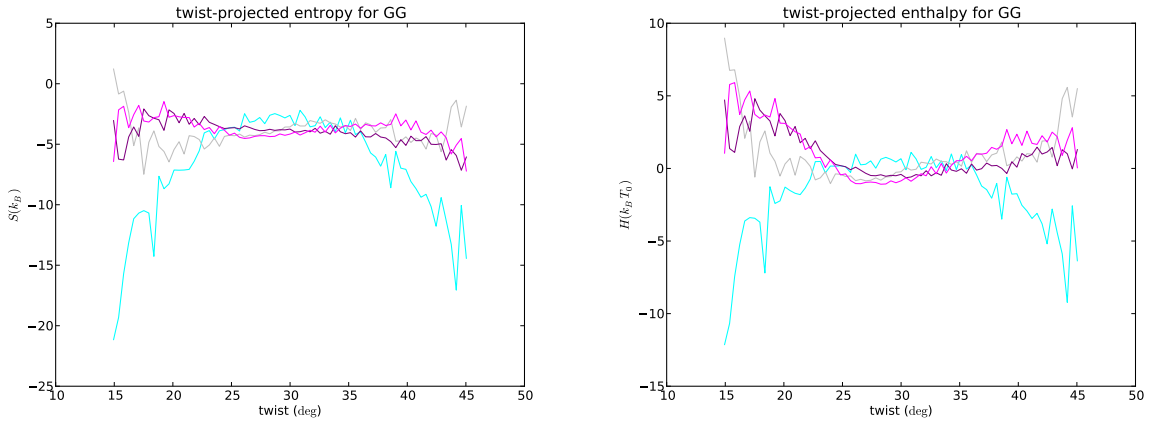


FIGURE 1.20 – (A) Pseudo-entropy computed from the projection on the twist axis, for the dinucleotide GG. Temperatures : 278K (cyan), 291.5K (gray), 312.5K (purple), 337.5K (magenta). (B) Pseudo-enthalpy (same colors)

### 1.7.2 Methods for the error analysis

Let us consider the successive values $\{x_i\}_{i=1,\ldots,N}$ of a quantity $x$, taken from a molecular dynamics trajectory, for instance the fluctuating position of a particle. We restrict the following derivation to a unidimensional system, and simply discuss the generalization to multidimensional systems. We address the problem of estimating the mean value and the variance of the distribution, and error estimated for these two quantities.

#### (a) Estimators

The quantity $x$ is assumed to fluctuate around a *mean value* $\mu = \langle x \rangle$, with a *variance* $\sigma^2 \equiv \sigma^2(x) = \langle x^2 \rangle - \langle x \rangle^2$. In the following, we use Greek letters $(\mu, \sigma)$ for the exact parameters of a distribution, and Latin letters $(m, s)$ for their estimators. We also abusively use the same symbol $\sigma^2$ for the variance *function* of a given distribution and for the particular *value* of the variance of $x$.

**Mean value** An expectation value for the mean value is given by the estimator :

$$m \equiv \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1.26}$$

The confidence interval (or error estimate) for the computed value of the mean is obtained from the square root of *variance of* $m$ :

$$\sigma^2(m) \equiv \langle m^2 \rangle - \langle m \rangle^2 \tag{1.27}$$

We introduce the *time correlation function* $C_{i,j}$, which for a system at equilibrium depends only on the difference between $i$ and $j$ (time invariance) :

$$C_{i,j} \equiv \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle = C_{i-j} \tag{1.28}$$

By plugging Eq. 1.26 into Eq. 1.27, we then get :

$$\sigma^2(m) = \frac{1}{N^2} \sum_{i,j=1,\dots,N} (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle)$$

$$= \frac{1}{N^2}[N \sum_i C_{i,i} + 2 \sum_{i<j} C_{i,j}]$$

$$\sigma^2(m) = \frac{1}{N}[C_0 + 2 \sum_{n=1}^{N-1} (1 - n/N) C_n] \tag{1.29}$$

where in the second line we have separated the terms where $i = j$. Note that the time correlation function at time 0 is just the variance : $C_0 = \sigma^2$.

In the case where the $x_i$'s are uncorrelated (the time between two measures is larger than the correlation time of the dynamic process), $C_i = 0$ for all $i \neq 0$ and the error on the estimation of the mean value reduces to :

$$\sigma^2(m) = \frac{\sigma^2}{N} \equiv \frac{\sigma^2(x)}{N} \tag{1.30}$$

$$\sigma(m) = \frac{\sigma}{\sqrt{N}} \tag{1.31}$$

**Variance** We proposed to estimate reliably not only the mean value of the coordinates, but also their variance $\sigma^2$, which is given by the estimator $s^2$ :

$$s^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{1.32}$$

Note that the sum runs over $N$ terms, and is divided by $N-1$ : this is necessary to get the *unbiased* estimation of the variance, with $\langle s^2 \rangle = \sigma^2$ :

$$\langle s^2 \rangle = \frac{1}{N-1} \sum_{i=1}^{N} \langle x_i^2 + \bar{x}^2 - 2 x_i \bar{x} \rangle = \frac{1}{N-1} \sum_{i=1}^{N} [\langle x_i^2 \rangle + \frac{1}{N^2} \left\langle \sum_{j,k} x_j x_k \right\rangle - \frac{2}{N} \left\langle \sum_j x_i x_j \right\rangle]$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} [\sigma^2 + \frac{1}{N^2} \sum_{j,k} \delta_{j,k} \sigma^2 - \frac{2}{N} \sum_j \delta_{i,j} \sigma^2] = \frac{1}{N-1} N \sigma^2 [1 + \frac{1}{N^2} N - \frac{2}{N}] = \sigma^2$$

This estimator will be used in the computation of the variance, as well as the confidence interval Eq. 1.30. In fact, here we propose to go even one step further, and estimate if the latter interval can be trusted ! This involves to compute the fourth moment of the distribution :

$$\sigma^4(x) \equiv \sigma^2(\sigma^2(x)) = \left( \left\langle \left( x^2 - \langle x \rangle^2 \right)^2 \right\rangle - \left\langle x^2 - \langle x \rangle^2 \right\rangle^2 \right) = (\langle (x - \langle x \rangle)^4 \rangle - \langle (x - \langle x \rangle)^2 \rangle^2) \tag{1.33}$$

Assuming that $x$ follows a Gaussian distribution, we have the relation $\langle x^4 \rangle = 3\sigma^4$. And :

$$\sigma^2(\sigma^2(x)) = 2\sigma^4 \tag{1.34}$$

From Eq. 1.30, *the expectation value for the error on the mean* is then given by :

$$\langle \sigma^2(m) \rangle = \frac{1}{N} \left( \sigma^2 \pm \sqrt{\frac{2\sigma^4}{N}} \right) = \frac{\sigma^2}{N} \left( 1 \pm \sqrt{\frac{2}{N}} \right) \tag{1.35}$$

### (b) The block averaging method

The previous formula Eq. 1.35 together with the estimator for the variance Eq. 1.32 allow us to compute reliable estimates for the error bars. However, they are valid only when the data-points are uncorrelated, which is not the case *a priori* for our time series. The *block averaging* method [Flyvbjerg and Petersen 1989] is an efficient way to tackle this problem, providing the correlation time together with a reliable estimate of the error. We first develop the method for the estimation of the *error on the mean value* of a sample, and then consider other quantities. Although the procedure is described here in terms of MD trajectories, it is also applicable to MC sampling paths.

**Mean value** The method consists in transforming iteratively the dataset $\{x_i\}_{i=1,\dots,N}$ into a new dataset :

$$\{x'_i = \frac{x_{2i} + x_{2i+1}}{2}\}_{i=1,\dots,N/2} \tag{1.36}$$

This operation conserves the estimate of the mean value : $m' \equiv \bar{x'} = m$. It can also be shown that the variance in the estimation of the mean is conserved : $\sigma^2(m') = \sigma^2(m)$ [Flyvbjerg and Petersen 1989].

Without any assumption on $\{x_i\}$, Eq. 1.29 shows that :

$$\sigma^2(m) \geq \frac{C_0}{N} \tag{1.37}$$

This is an equality only when the $\{x_i\}$ are uncorrelated, in which case the remaining terms vanish. Conversely, if the data-points are correlated, $C_0/N$ is an underestimation of the error : if $n_c > 1$ is the correlation time, we have $\sigma^2(m) \simeq \frac{C_0}{N/n_c} > \frac{C_0}{N}$.

Upon iterative application of the transformation 1.36, the dataset decreases in size, and the data-points become less correlated. Thus, if at each step we compute the value $\frac{C_0}{N}$ using the estimator 1.32, the values increase until the reduced dataset has become uncorrelated, and the value becomes invariant by the transformation. For a trajectory of total size $N_t$ and a block size $n_b$, we define the computed quantity as the *block-estimated error*, given by :

$$\tilde{\sigma_m}^2(N, n_b) \equiv \frac{s^2(x_i^{n_b})}{N/n_b} \left(1 \pm \sqrt{\frac{2}{N/n_b}}\right) \tag{1.38}$$

The curve gives the following information simultaneously :

- if the constant regime is not reached for $n_b \to N_t$, the sampling is insufficient, and we only have a lower bound on the error
- the transition between the two regimes provides a characteristic time of the sampled quantity, which will be discussed further in the case of the harmonic oscillator
- the value of the fixed point (with its confidence interval) is a reliable estimate of the error on the mean value

**Variance** The same kind of procedure can be used to estimate the errors on other quantities, as will be the case when treating the data. Here, for a unidimensional process, we develop the example of the variance error. *A priori*, one may consider two possible methods :

1. we operate on the same dataset $\{x\}$, by computing the analogue of Eq. 1.38 for the variance. However, this involves the computation and unbiased estimation of higher moments of the distribution (the eight moment in this case !)
2. we construct a new dataset $\{v_i = x_i^2 - \bar{x^2}\}$, for which *the error on the mean value* is precisely *the error on the variance of x*, as shown below. We can then simply re-use Eq. 1.38 on the new dataset.

In the following calculation, we show that the squared error on $\bar{v}$, $\sigma^2(\bar{v}) = \sigma^2(v)/N$, is indeed equal to the *error on the variance of x*, $\sigma^2(\sigma^2(x))/N$. $v$ has a null average value :

$$\langle v \rangle = \langle x_i^2 \rangle - \left\langle \bar{x^2} \right\rangle = 0 \tag{1.39}$$

So, its variance reduces to the squared term :

$$\begin{aligned} \langle v_i^2 \rangle = \left\langle (x_i^2 - \bar{x^2})^2 \right\rangle &= \left\langle x_i^4 + (1/N\sum_i x_i^2)(1/N\sum_j x_j^2) - 2/N x_i \sum_j x_j^2 \right\rangle \\ &= \left\langle x_i^4 - 1/N\sum_j x_i^2 x_j^2 \right\rangle = (1-1/N)(\langle x_i^4 \rangle - \langle x_i^2 \rangle) \end{aligned} \tag{1.40}$$

Finally, the error on the $v$ mean value is given by :

$$\langle \sigma^2(\bar{v}) \rangle = \frac{1}{N} \langle s^2(v) \rangle = \frac{1}{N(N-1)} \langle v_i^2 \rangle = \frac{1}{N}\sigma^4(x) = \frac{1}{N}\sigma^2(\sigma^2(x)) \tag{1.41}$$

*i.e.* the error in estimating the mean value of $v$ is equal to the error in the variance of $x$.

In the analysis of the MD trajectories, we extend this method to the moments of the multidimensional trajectory (covariance), but also to more indirect quantities, like the stiffness. In the latter case, the error estimate does not rely on rigorous calculations as here, but rather on the quite natural idea that if we cut the trajectory into slices, the variance of any quantity computed independently on these slices provides an estimate of the error.

### 1.7.3 Harmonic oscillator model : exact results for the error estimates

#### (a) Simple damped harmonic oscillator

The previous calculations were applicable to any trajectory, without any assumption on the correlation function, except that the trajectory is sufficiently long to sample independent (uncorrelated) points. Here, we consider a very simple model : the damped harmonic oscillator. In this case the correlation function is a simple decreasing exponential with a characteristic time $\tau$. We are able to develop exact analytical calculations, and relate them to the values computed previously, allowing a better understanding of the procedure. To simplify the calculations, we consider here an oscillator centered on zero.

The Langevin equation for an over-damped harmonic oscillator centered at 0 is given by [Barrat and Hansen 2003] :

$$0 = -kx - \zeta\dot{x} - f(t) \tag{1.42}$$

where $\zeta$ is a friction coefficient and $f(t)$ is a stochastic force characterized by :

$$\langle f(t) \rangle = 0 \ ; \ \langle f(t) f(t') \rangle = 2\zeta k_B T \delta(t-t') \tag{1.43}$$

The solution is given by :

$$x(t) = x(0)\exp(-t/\tau) + \sigma^{-1}\int_{t'=0}^{t}\exp((t-t')/\tau)f(t')dt' \tag{1.44}$$

where $\tau = \sigma/k$ is the correlation time, with :

$$\langle x(t)x(0)\rangle = \sigma^2\exp(-t/\tau) \tag{1.45}$$

**Mean value** If the mean value is computed along a trajectory of finite length $T$, there is an expected error :

$$\sigma^2(m_T) = \left\langle\left(\frac{1}{T}\int_0^T x(t)dt\right)^2\right\rangle = \frac{2}{T^2}\int_0^T dt_1\int_{t_1}^T dt_2\,\langle x(t1)x(t2)\rangle = \frac{2}{T}\int_0^T du\,\langle x(0)x(u)\rangle\,(T-u)du \tag{1.46}$$

In the last line, we used the fact that the correlation function depends only on the time difference to change variables. Replacing its value by Eq. 1.45, we get :

$$\sigma^2(m_T) = \frac{2\sigma^2\tau}{T}\left(1-\frac{\tau}{T}(1-e^{-T/\tau})\right) \simeq \frac{2\sigma^2\tau}{T}\left(1-e^{-T/2\tau}\right) \tag{1.47}$$

Notice that this is just the continuous version of Eq. 1.29. The most important case is when $T \gg \tau$, in which case we can compare the value to the discrete case Eq. 1.35 to find the number of independent points $N_i$ in a MD trajectory of an harmonic oscillator :

$$\sigma^2(m_{T\gg\tau}) \simeq \frac{2\tau}{T}\sigma^2 \qquad N_i(T,\tau) = \frac{T}{2\tau} \tag{1.48}$$

In the opposite regime $T \ll \tau$, the error decreases exponentially, with a characteristic time $\tau/2$. From this equation, it is possible to compute the value of the "block function" $\tilde{\sigma_m}^2(N,n_b)$, with $T = N\,\Delta t$ and $t_b = n_b\,\Delta t$, where $\Delta t$ is the timestep of the trajectory. For a block time $t_b$, $\tilde{s_m}^2(T,t_b)$, is given by the observed variance on the block mean values, $\sigma^2(m_{t_b})$, divided by the number of blocks in the trajectory, $T/t_b$. Using Eq. 1.38, we get :

$$\tilde{s_m}^2(T,t_b) \simeq \frac{2\sigma^2\tau}{T}\left(1-e^{-t/2\tau}\right) \tag{1.49}$$

**Variance** The error on the variance can be estimated analytically in the same way. In analogy with Eq. 1.46, it is given by :

$$\sigma^2(m(\sigma^2(x))_T) = \left\langle\left(\frac{1}{T}\int_0^T x^2(t)dt - \left\langle x^2\right\rangle\right)^2\right\rangle = \sigma^2(\mu(v)_T) \tag{1.50}$$

*i.e.* it is equal to the error in the estimation of the mean of the variable $v$ defined as :

$$v = x^2 - \langle x\rangle^2 \tag{1.51}$$

Note the correspondence with the definition of the $v_i$ in the discrete case, Eq. 1.40. The trick is therefore to re-use the results computed for the mean value, but this time for the distribution $v$. From equation 1.46, this time for $v$ :

$$\sigma^2(m(\sigma^2(x))_T) = \frac{2}{T^2}\int_0^T dt\ \langle v(t)v(0)\rangle\,(T-t) \tag{1.52}$$

We thus need to compute the correlation function of $v$ :

$$\langle v(t)v(0)\rangle = \left\langle (x(t)^2 - \left\langle x^2\right\rangle)(x(0)^2 - \left\langle x^2\right\rangle)\right\rangle = \left\langle x(t)^2x(0)^2\right\rangle - \left\langle x^2\right\rangle^2 \tag{1.53}$$

To compute the correlation function, we square the solution of the Langevin equation, Eq. 1.44 :

$$x(t)^2x(0)^2 = x(0)^4e^{-2t/\tau} + \frac{2x(0)^3}{\zeta}e^{-t/\tau}\int_0^t e^{-(t-t')/\tau}f(t')dt' + \frac{x^2(0)}{\zeta^2}\int_0^t\int_0^t du\,dv\,e^{-(2t-u-v)/\tau}f(u)f(v) \tag{1.54}$$

Taking the average of this expression, from the properties of the noise function, Eq. 1.43, the second term is zero and the third term reduces to a single integral :

$$\begin{aligned}\left\langle x(t)^2x(0)^2\right\rangle &= \left\langle x_0^4\right\rangle e^{-2t/\tau} + \frac{1}{\zeta^2}\int_0^t du\,e^{-(2t-2u)/\tau}2k_BT\zeta \\ &= 3\sigma^4e^{-2t/\tau} + \frac{2k_BT\sigma^2}{\zeta}\tau(1-e^{-2t/\tau}) = (1+2e^{-2t/\tau})\,\sigma^4\end{aligned} \tag{1.55}$$

Finally, plugging Eqs. 1.53 and 1.55 into 1.52, we get :

$$\sigma^2(m(\sigma^2(x))_T) = \frac{2}{T^2}\int_0^T dt\,2\sigma^4e^{-2t/\tau}(T-t) = \frac{2\tau}{T}\sigma^4\left(1-\frac{\tau}{2T}(1-e^{-2T/\tau})\right) \tag{1.56}$$

In the case where the sampling is sufficient, $T \gg \tau$, this simplifies into

$$\sigma^2(m(\sigma^2(x))_{T\gg\tau}) \simeq \frac{2\tau}{T}\sigma^4 \tag{1.57}$$

By comparing this result to Eq. 1.41, we see that the error on the variance behaves very similarly to that of the mean value. In particular, the number of independent points in the trajectory is the same. However, the block function analogous to Eq. 1.49 is now given by :

$$\tilde{s_v}^2(T,t_b) \simeq \frac{2\sigma^4\tau}{T}\left(1-e^{-t/\tau}\right) \tag{1.58}$$

While the characteristic time of the mean block curve is $\tau/2$, for the variance it is $\tau$.

### (b) Superposition of two independent harmonic oscillators

In all the previous developments, the trajectory involved only a single correlation time. However, in the MD trajectories that we treat, there is no reason that it should be the case. In particular, we consider 6-dimensional trajectories, with at least a correlation time for each dof. Correlations between these dof will mix the different times. To better understand the possible consequences of this issue, we treat here the simplest possible model of a process with two correlation times. Here we restrict to the estimation of the *mean value* for simplicity.

The variable $z = x + y$ is the sum of two independent harmonic oscillators, characterized by variances $\sigma_x, \sigma_y$ and correlation times $\tau_x, \tau_y$. The variance is given by : $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$. The correlation function of $z$ is then :

$$\langle z(0)z(t)\rangle = \sigma_x^2 e^{-t/\tau_x} + \sigma_y^2 e^{-t/\tau_y} \tag{1.59}$$

Interestingly, this function can have different behaviors, depending on the relative values of the variances and the correlation times. For instance, if $\sigma_x \gg \sigma_y$ while $\tau_x \ll \tau_y$, the static variance of $z$ will be dominated by $x$, while the decay of the time correlation will be dominated by $y$.

This has a direct consequence for the the error estimates, as can be computed from Eq. 1.47 :

$$\sigma^2(m_T(z)) = \frac{2}{T^2}\int_0^T \langle z(0)z(t)\rangle\,(T-t)dt \simeq \frac{2\sigma_x^2\tau_x}{T}\left(1-e^{-T/2\tau_x}\right) + \frac{2\sigma_y^2\tau_y}{T}\left(1-e^{-T/2\tau_y}\right) \tag{1.60}$$

This means that the "block curve" is the sum of two exponential functions, with decay times $\tau_x/2$, $\tau_y/2$. If they are very different, the slowest one only will dominate the decay of the curve for long times, as for the correlation function. The typical value of the error at long times is given by :

$$\sigma^2(m_{T\gg\tau_x,\tau_y}(z)) = \frac{2}{T}(\sigma_x^2\tau_x + \sigma_y^2\tau_y) \tag{1.61}$$

Note that this is exactly the same result as for a single harmonic oscillator, Eq. 1.48, where we introduce the total correlation time $\tau_z$ :

$$\sigma^2(m_{T\gg\tau_z}(z)) = \frac{2\tau_z}{T}\sigma^2 \quad ; \quad \tau_z \equiv \frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2}\tau_x + \frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}\tau_y \tag{1.62}$$

This definition is consistent with the standard defition of the correlation time : $\tau_z = \int_0^\infty dt\,\langle z(0)z(t)\rangle \,/\, \langle z(0)^2\rangle$. This total correlation time, which governs the error estimate, may be different from $\tau_x$ and $\tau_y$. As an interesting example, let us consider the academic case where $\sigma_x^2 = 1$, $\sigma_y^2 = 0.1$, $\tau_x = 0.1$, and $\tau_y = 1$. It corresponds to the situation, where a rapid motion $x$ explores the large values, and thus dominates the static variance $\sigma_z^2$, but where the slower motion dominates the correlation function for longer times (see Section 1.3.1).

In the error function Eq. 1.60, both terms yield a contribution of similar magnitude, however the $x$ term vanishes much more rapidly. As a consequence, the $y$ term only will be visible in the "block curve" (see Eq. 1.49), with a decay time of $\tau_y/2 = 0.5$. On the other hand, the total correlation time $\tau_z$ which dominates the error is approximately equal to 0.2, *i.e.* much smaller. Interestingly, this very situation is commonly found in the block curves computed on the MD trajectories (see Section 1.3.1).

As an example, in Fig. 1.21, we compare the block curve of a given parameter in the real trajectory (left) and on the artificial trajectory generated with from the same effective correlation time. By the latter choice of construction, the error is similar, and yet the appearance of the curves is very different. The reason is that the real data contains a mix of rapid ($\sim$ 50 ps) and slow ($\sim$ 1000 ps) processes, and while the latter dominates the appearance of the curve, the error is dominated by the former, which is quite a surprising result !

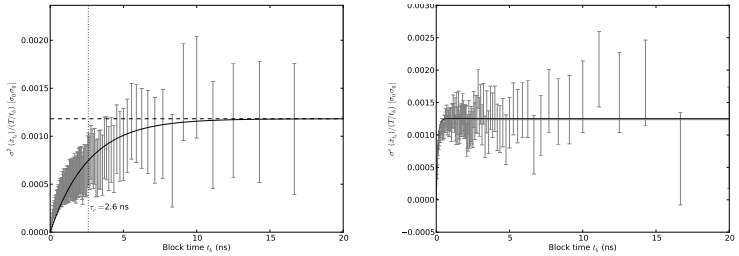


FIGURE 1.21 – Block errors for the mean rise of AA at 300K, as computed on the real data (left) and artificial data generated with the same effective correlation times. By construction, the two curves provide the same squared error ($\sim 0.0012$ in the chosen units). However, the curve of real trajectory is dominated by a slow process ($\tau = 2.6$ns), which is absent of the artificial one. This is an illustration of the specific situation where a rapid motion dominates the variance ($\tau \simeq 50$ps), and therefore also the error.

### 1.7.4 Validation of the error analysis on artificial data

#### (a) Generation of a unidimensional harmonic oscillator trajectory

In the previous paragraphs, we computed the error estimate in the case of an harmonic oscillator, which will be used to estimate the correlation times present in the MD data. To validate the analysis procedure, we will generate artificial trajectories that "mimic" the real ones. Here, we give the stochastic equations for the construction of a trajectory of N points, centered on 0, a reduced time $t_r = \Delta t/\tau$ between two points and a variance $\sigma^2$. The latter quantities can be related to the parameters of the previous paragraph : $\tau = \zeta/k$ and $\sigma^2 = k_B T/k$. The equations are given from the development of Eq. 1.44 in the case of a small $\Delta t$.

The initial value is taken randomly from a normal distribution centered on zero and of variance $\sigma^2$ :

$$x_0 = \mathcal{N}(0, \sigma^2) \tag{1.63}$$

The increment is given by :

$$x_{i+1} = x_i e^{-t_r} + \mathcal{N}(0, (1 - e^{-2t_r})\sigma^2) \tag{1.64}$$

#### (b) Error estimates on artificial datasets

To validate our analysis method, we generate artificial data that mimic the properties of the real one, and where we repeat the procedure described previously. Doing so, (i) we know the "exact" model with which we have constructed the trajectory, and thus we can compare the error estimates with "real errors" and (ii) there are no systematic errors as in the real data : we can therefore separate this source of errors from the statistical problems. For this part, we used the data from the step parameters of the sequence "AA", where we have the weakest statistics.

Section (a) gives the stochastic equations for generating the trajectory of a damped harmonic oscillator in 1D, with a prescribed trajectory time T, correlation time $\tau_c$ and variance $\sigma^2$. Starting from the real 6-dimensional trajectory we wish to mimic, the procedure is the following :

- We estimate the covariance matrix with Eq. 1.18. This matrix is an *estimation* of the real correlations, characteristic of the underlying physical process. In the artificial data, we use this estimate as the "real" covariance matrix.
- The eigenvectors $\{\vec{v}_i\}_{i=1,\ldots,6}$ of this covariance matrix form a base, where the datapoints separate into 6 uncorrelated unidimensional trajectories. Using the transformation matrix

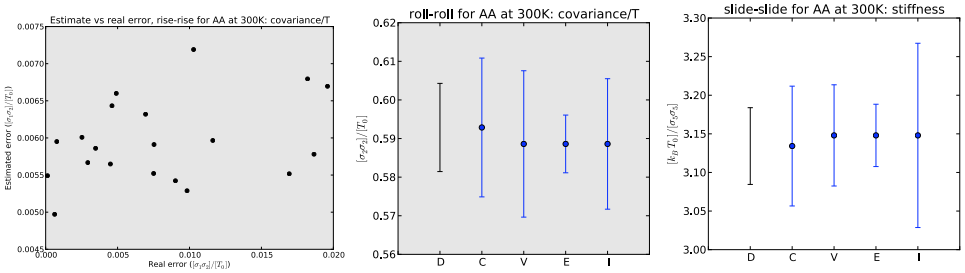


FIGURE 1.22 – The artificial data allows to test the validity of the error analysis. 20 artificial trajectories, which mimic a real trajectory (see text), and the analysis was carried on each of them. **(left)** Comparison of estimated errors with the real errors : they are in the same range, even if sometimes (right part of the figure) the larger errors are underestimated. **(center)** Comparison of different computed values and estimated errors for a covariance element : D- estimated value and error for the real data, which serve to parametrize the artificial trajectory ; C- Expected value and error, as provided by the analytical calculations on the employed model ; V- Average and width of the values computed in the 20 trajectories : the expected distribution is well reproduced ; E- Average value of the estimated error, which is indeed underestimated. The left hand-side figure shows that for some trajectories, the large errors were not accurately detected ; I- Average value of the error, as estimated after inverting the estimated errors of the stiffness by a Monte Carlo procedure (see text). Here the hypothesis of independent errors results in a large error bar, which partly compensates for the possibly small estimated stiffness errors. **(right)** Same for the stiffness, with same features. Here the inverted bar is even larger.

$\underline{V}_{ij} = (v_i)_j$, we generate this "eigentrajectory" $\vec{q}'_i = \underline{V}^t \vec{q}_i$. By construction, in this new base the covariance matrix is a diagonal matrix, where the coefficients are the eigenvalues $\lambda_i$ of the original one.

- In this base, we operate the block averaging method separately in the 6 dimensions, which gives the 6 effective correlation times associated to the uncorrelated processes. Using these times and the corresponding variances $\lambda_i$, we can generate 6 unidimensional artificial trajectories, $\vec{q}'_{art,i}$, which have similar statistics as the original data in the eigenbase.
- Finally, we use the matrix $\underline{V}$ to map the artificial datapoints back to the standard base, $\vec{q}_{art,i} = \underline{V}\vec{q}'_{art,i}$.

As a result of this process, the different degrees of freedom are correlated in a similar way as the real data, and the statistical properties are similar.

We made some preliminary tests in 1D, where we avoid the possibly subtle effects related to the mixing of different correlation times. By generating multiple replicates of a trajectory with similar statistics as the real data, we can test the rapidity of convergence of the error estimates. When fitting the "block curves", Fig. 1.7, we noticed that for 5ns blocks (*i.e.* 1/20 of the total trajectory), the error estimates exhibit $\sim 30\%$ relative fluctuations. Here, we made an analogous calculation, by computing the dispersion of the estimated variance over 20 complete trajectories. Repeating this operation shows that the dispersion is indeed of $\sim 35\%$, confirming that the lack of precision in the determination of error bars is due to the limited available statistics and not to a error in the procedure.

In a second step, we generated and analyzed 20 replicas of same trajectory mimicking the data, where we took the parameters from the step datapoints for AA at 300K, and we systematically compare estimated and real errors. Fig. 1.22(left) shows that for most cases (left half of the plot), the values are compatible, but sometimes more important errors are indeed largely underestimated (right half). There is no particular correlation between the two. Then, we can

test what happens when inverting the matrix : *i.e.* if the hypothesis of independent errors for the different elements of the matrix results in a large overestimation of the errors (see previous section). For the roll-roll covariance, we therefore compare (i) the deviation of computed values in the different trajectories (quoted V), (ii) the mean value of the estimated error (quoted E) and (iii) the mean error, as computed from inverting the stiffness, with the MC procedure described in the previous section (quoted I). These are the three bars at the right of Fig. 1.22 (center) and (right) : the plot shows that the bars are of the same order, but the estimated error is indeed smaller, in the range of uncertainty mentioned earlier. However, this effect is somehow corrected by the hypothesis of independent errors for the matrix inversion : the final result of the model, which can be directly compared to the covariance seen in the data, is the "I" value.

Finally, we generated a whole T-dependent set of trajectories following the fitted model of the stiffness, and operated the complete analysis on it. Without the outlying points (at 273K in Fig. 1.23, upper panels), the analysis is consistent (lower panels).

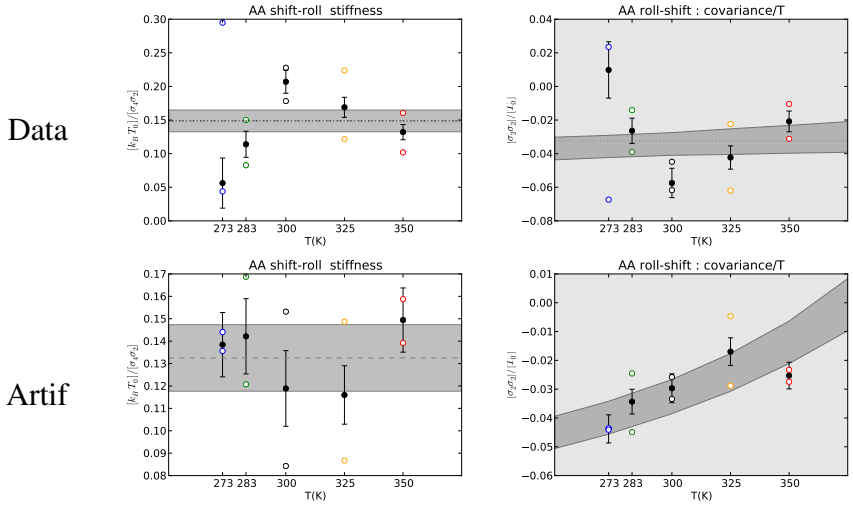


FIGURE 1.23 – Comparison of the analysis procedure in the real (upper panels) and corresponding artificially generated data (lower panels) for a matrix element where the data is of poor quality : fit of the stiffness (left) and resulting model for the covariance (right). Here the effect of temperature is very limited, and the error bars seem slightly underestimated, but in the large range mentioned earlier. This is not true for the point at 273K, which is outlying : this systematic error is a consequence of a bad equilibration, as demonstrated by the points corresponding to the two positions of the sequence (colored dots), and is not accurately detected by our procedure. In the artificial data where these problematic features have been eliminated, the procedure is robust.

### 1.7.5 Detailed results : internal bp parameters

| buckle | propel | opening | shear | stretch | stagger |
|---|---|---|---|---|---|
| 1 | 6 | 3 | 5 | 3 | 7 |
| 6 | 3 | 5 | 3 | 6 | 3 |
| 3 | 5 | 11 | 6 | 5 | 4 |
| 5 | 3 | 6 | 11 | 11 | 6 |
| 3 | 6 | 5 | 11 | 11 | 9 |
| 7 | 3 | 4 | 6 | 9 | 11 |

TABLE 1.8 – Number of dinucleotides where a significant entropic contribution was detected in the data, for each element of the 6x6 stiffness matrix

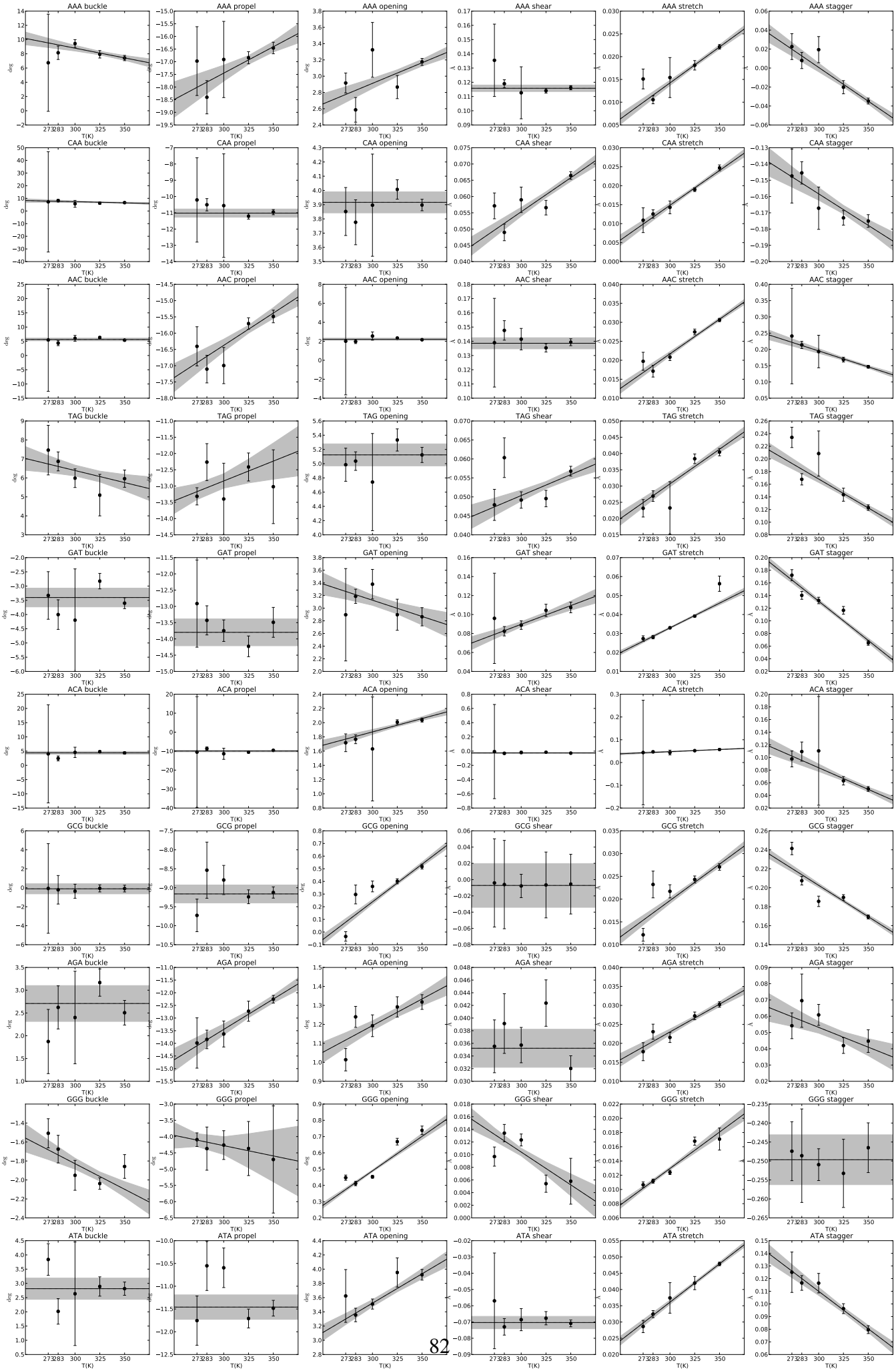




FIGURE 1.24 – Evolution of the equilibrium values of the intra parameters, for all sequences (rows).

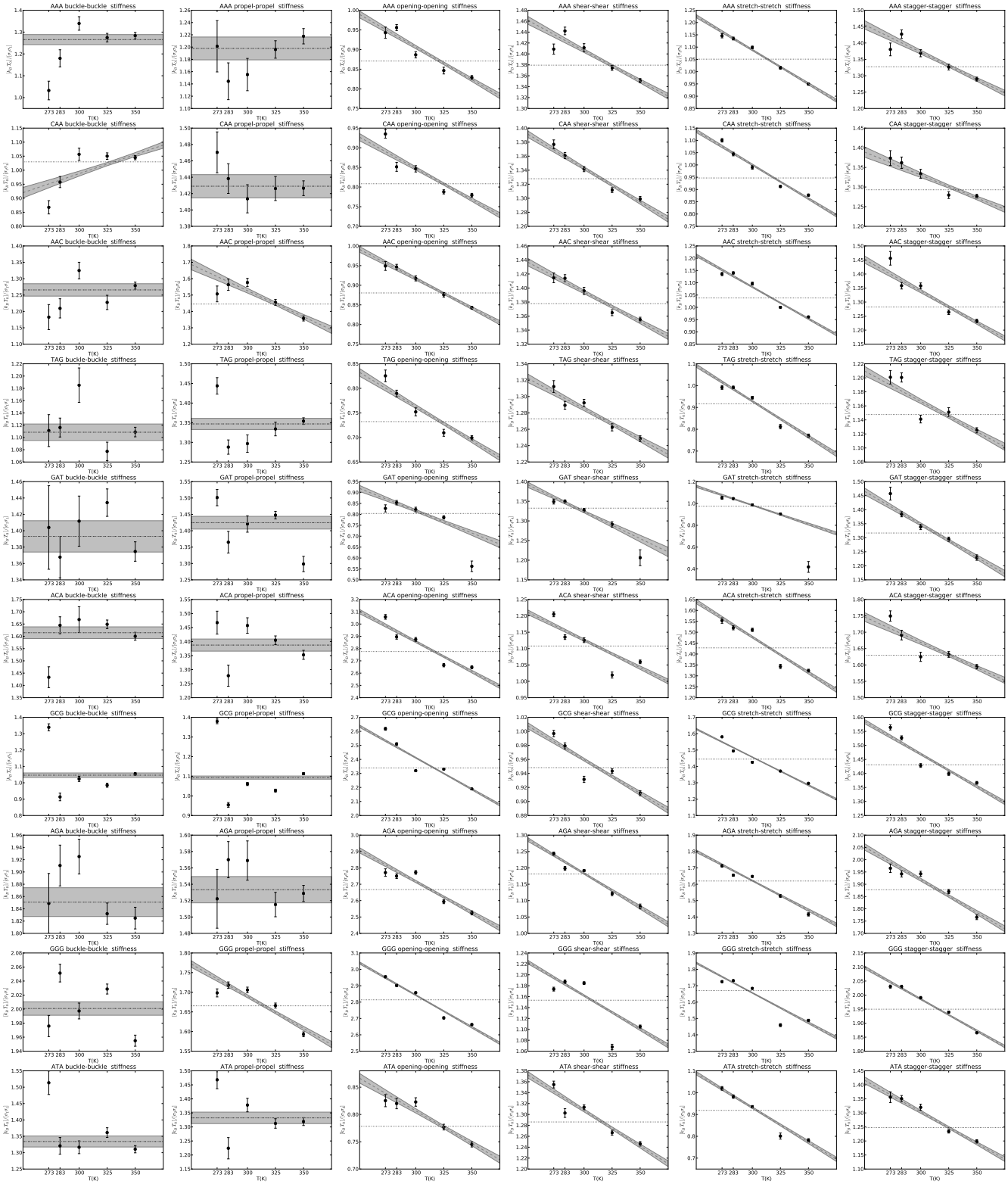


FIGURE 1.25 – Regression of the diagonal stiffness elements of the intra parameters, for all sequences (rows).

### 1.7.6 Detailed results : bp-step parameters

| tilt | roll | twist | shift | slide | rise |
|---|---|---|---|---|---|
| 9 | 3 | 4 | 6 | 4 | 5 |
| 3 | 9 | 3 | 2 | 2 | 5 |
| 4 | 3 | 3 | 5 | 7 | 5 |
| 6 | 2 | 5 | 3 | 5 | 5 |
| 4 | 2 | 7 | 5 | 7 | 5 |
| 5 | 5 | 5 | 5 | 5 | 11 |

TABLE 1.9 – Number of successful dinucleotides for each element of the stiffness matrix

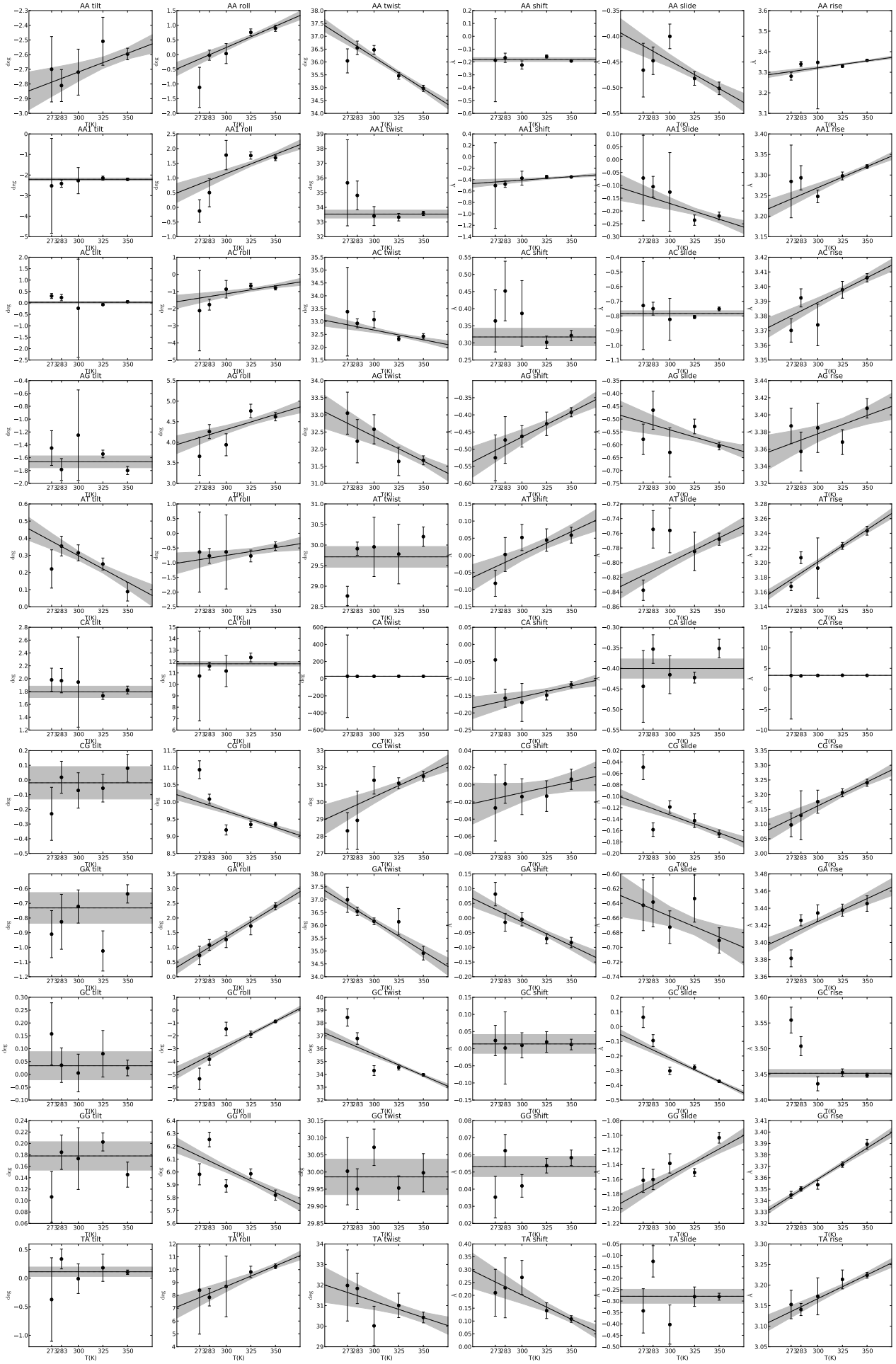


FIGURE 1.26 – Evolution of the equilibrium values of the step parameters, for all sequences (rows). The only parameter where a systematic effect can be identified is rise (last column). The spontaneous twist exhibits little temperature dependence, indicating no spontaneous supercoiling.

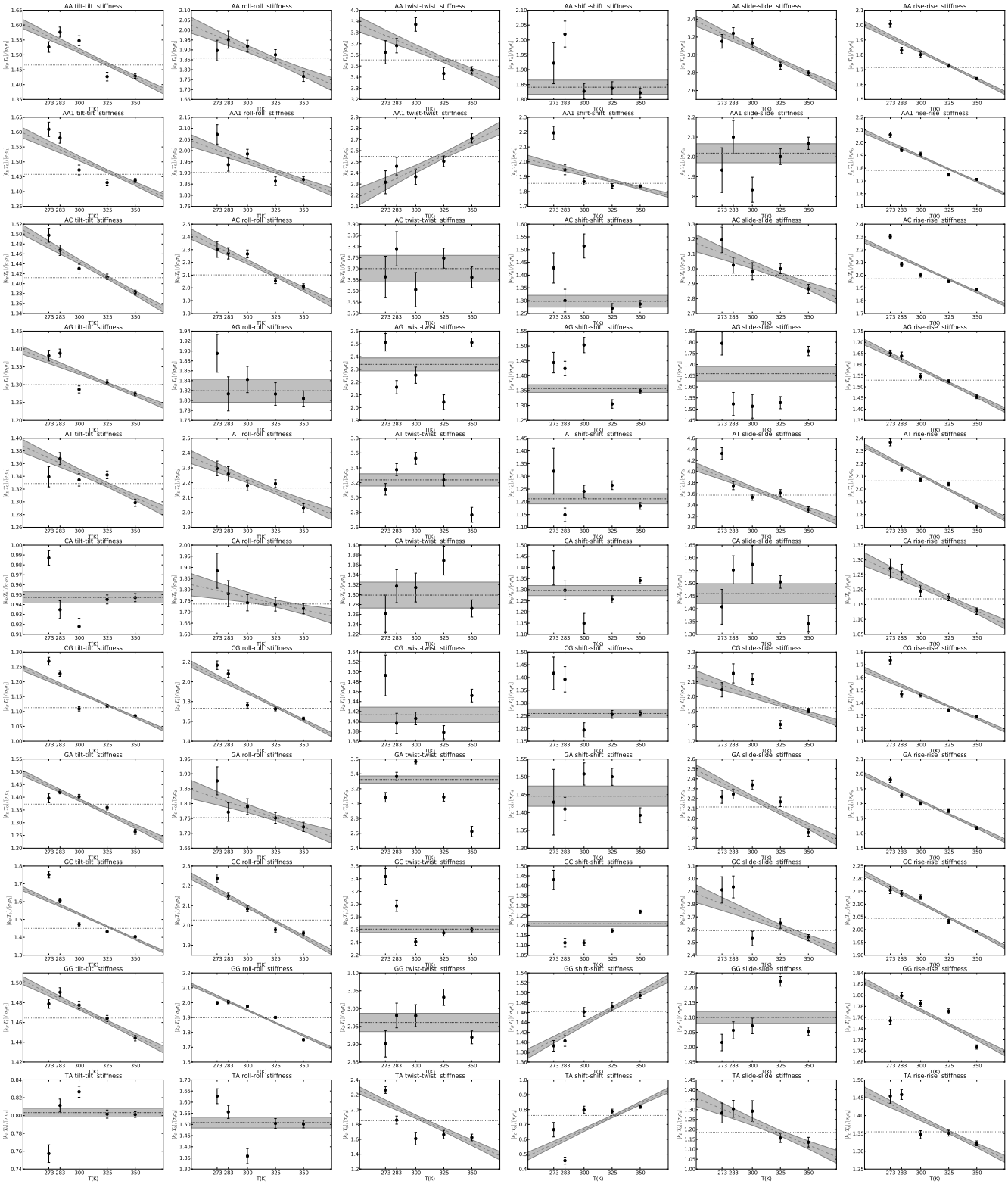


FIGURE 1.27 – Regression of the diagonal stiffness elements of the step parameters, for all sequences (rows).

# Chapitre 2

# DNA mechanics in the nucleosome

*In this chapter, we focus on the mechanics of the histone-DNA interactions in the Nucleosome Core Particle (NCP). We analyze crystallographic and MD-derived nucleosome structural models, and discuss the consequences of the inferred forces for the mechanics of the nucleosome. By lack of time, only the first results are discussed.*

## Introduction

The wrapping of DNA in the nucleosome is the basic mechanism of its compaction in the nucleus. At first sight, this mechanism is very surprising for the physicist, due to the severe bending involved. The required amount of energy is mediated by the relatively short-ranged electrostatic interactions between the histones and the bound DNA.

The knowledge of these interactions, and thus of the physics of the nucleosomes, has greatly benefited from the high-resolution structures obtained by X-ray crystallography in the last 15 years. In particular, the precise characterization of 14 discrete "anchor points", where positively charged amino-acids (AA) come into close contact with the DNA phosphates and even into the minor grooves, has strengthen the idea that the interactions are localized at these sites. But structural observations do not provide a quantitative model for the interactions. For this reason, the mechanical properties and the dynamics of the nucleosome were mainly studied with two models.

On the one hand, it is possible to describe explicitly all the atoms of the system, and let it evolve in Molecular Dynamics (MD) trajectories. This is the method of choice for problems where some precision is required ; its main drawback is the prohibitive computing time. Let us mention that the wrapped DNA alone contains around 6000 non-hydrogen atoms, without the histones and the solvent : the numerical complexity then restricts the studies to a limited trajectory time of $\sim$ 10 ns, incompatible with large-scale events. As an example, nucleosome breathing is estimated to have a typical time of $\sim$ 50 ms [Li et al. 2005].

The study of events extending on the lengthscale of the whole nucleosome thus involves the use of a coarse-grained model. Because of the lack of knowledge previously mentioned, the only existing coarse-grained models of the histone-DNA interactions are minimalistic : the interaction energy is typically taken proportional to the DNA adsorbed length [Kulic and Schiessel 2003b, Biswas et al. 2012]. Such models efficiently describe the most generic effects of nucleosome mechanics. They are also the most straightforward choice for the simulations of

polynucleosomes [Wedemann and Langowski 2002].

Between these two descriptions, there is a gap. For many processes within the nucleosome, there is a lack of an intermediate coarse-grained model, which would describe the mechanics of the DNA within the nucleosome in some detail, and in particular would take into account the sequence. A simple but important example is the prediction of the sequence-dependent energetic cost of nucleosome formation. Because of the sequence-dependent mechanical properties of DNA, each sequence has a different affinity with respect to nucleosome formation. These affinities have been measured for a group of sequences [Lowary and Widom 1998, Battistini et al. 2012] and empirical tables of sequence-dependent "scores" have been computed of such experiments. An important objective of any realistic model of the nucleosome at the considered scale is the prediction of these values. Other examples of relevant processes include the simulation of unwrapping events [Voltz et al. 2012], and the diffusion of defects [Kulic and Schiessel 2003a].

Such a nanoscale model of the nucleosome implies the description of the sequence-dependent DNA mechanical energy, and the DNA-histones interaction energy. For the former part, available models exist, in particular the rigid base-pair (rbp) model, which has been presented in Chapter 0

In contrast, there is a lack of a corresponding validated potential for the histone-DNA interactions, in this case acting on the DNA base-pair (bp). As a consequence, the mechanics-based estimations of sequence-dependent nucleosome association free energies are computed by threading the sequences onto a fixed NCP structural model, either approximated by a superhelix [Anselmi et al. 2000], or based on one or several crystal structures [Deniz et al. 2011]. This is unsatisfactory on the theoretical point of view, as different sequences can be expected to accommodate the structural constraints of nucleosome association by different conformations. It also certainly contributes to the limited results of those estimations. In this study, we address the extraction of this potential from available high-resolution data on the nucleosome.

To our knowledge, the construction of a corresponding nanoscale mechanical model has been addressed in a single study only [Morozov et al. 2009]. With the same rigid-base pair model of DNA (and different parameters), Morozov et *al.* have built a phenomenological potential in which the DNA bp experience an elastic translational potential directed toward the superhelical path of DNA, as determined from a fit of the NCP147 crystal structure (PDB ID 1kx5). Interestingly, the structures obtained after relaxing the DNA in this potential exhibited some of the periodic structural features observed in the crystals, such as twist and roll oscillations. Note that the same kind of features is also detectable by relaxing the DNA without any external potential, but keeping the first and last bp fixed in their crystal conformation [Becker and Everaers 2009b]. Predicted sequence-dependent binding affinities were found to correlate with the experimental values.

Here, we propose a slightly different approach. We notice that the crystallographic structural models are quite distant from an idea superhelix, and there is some evidence that local features yield an important contribution in the DNA mechanical energy [Richmond and Davey 2003]. This is especially true at the 14 anchor points, where the bp are strongly deformed, and where the forces are likely to concentrate : the hypothesis of a force field homogeneously distributed along the molecule would then be inaccurate. The problem is that such features are difficult to estimate from the structures alone, because the relation between the state of deformation and the externally applied force is nontrivial and often surprising in an elastic object : for instance, one can have a strong deformation without any local force, because of the propagation of remotely

applied forces [Becker and Everaers 2009a]. The nanomechanical analysis allows to avoid this confusion, and to infer the external forces responsible for a given state of deformation : see Chapter 0. When applied on the nucleosome, one finds that the forces are indeed concentrated at the anchor points, exhibiting regular patterns [Becker and Everaers 2009a]..

In the past years, a growing number of nucleosomal structures have been crystallized. In addition to a series of crystals containing sequences derived from the original "NCP147" human $\alpha$-satellite sequence [Luger et al. 1997], three crystals have been recently resolved with the 601 positioning sequence [Makde et al. 2010, Vasudevan et al. 2010]. There is no doubt that this number will still increase in the future, giving increasing information on the details of the interactions [1]. Concomitantly, all-atomic MD trajectories of an entire nucleosome were found to be stable, at least on the limited accessible time, and may potentially constitute another source of information. Therefore, instead of constructing a potential from ideal models, we address the problem of *extracting* a coarse-grained nucleosomal potential from high-resolution structures. Our attempt is described in the following sections :

1. **Harmonic model of the histone-DNA interactions :** We describe the model of a harmonic interaction between the histones and the DNA bp and we develop the principle of our method to extract its parameters from a set of structures.
2. **Structural dataset :** We describe the available set of crystallographic structural models and all-atomic MD snapshots, and we discuss its relevance wrt the proposed analysis by comparing the different structures
3. **Nanomechanical analysis :** We operate the nanomechanical analysis on the dataset, and we discuss the estimated values of the forces.
4. **Extraction of the potential :** We operate the nanomechanical analysis on the dataset, and we discuss the estimated values of the forces.
5. **Discussion :** As the reader may notice, this chapter could not be completed by lack of time, and because the data exhibited a surprising behavior, which is delicate to interpret. In this section, we suggest some ideas in this regard, and some possibilities of future developments.

## 2.1 Method : extraction of a harmonic nucleosome potential at the base-pair level

### 2.1.1 Harmonic nucleosome potential and sequence-dependence

The extraction of a coarse-grained free energy function out of a set of atomic structures has been successfully achieved in the past. In particular, a parameter set for the equilibrium values and stiffness constants of (naked) DNA at the base-pair level has been obtained from the analysis of a large number of crystallographic structures of DNA and DNA-protein complexes [Olson et al. 1998]. In that approach, the underlying assumption is that the observed distribution of structures follows the equipartition of energy at some effective temperature : see Chapter 0.

In the NCP, the problem is slightly more complex. The potential experienced by a DNA bp is the sum of two contributions : the internal DNA elastic energy, and the histone-bp interactions. An important feature exhibited by the crystallographic structures is that the proteins make

1. Actually, new structures have already been published since we began this study, which are not included in our database, see for instance [Chua et al. 2012]

very few molecular contacts with the DNA bases, and therefore it has been hypothesized that the histone-DNA interactions are essentially sequence-independent [Olson and Zhurkin 2011]. The sequence preferences of the nucleosome would thus be the result of an indirect readout mechanism, *i.e.* mediated by the internal DNA elasticity.

The analysis of a "structural ensemble" of NCP structures can only give information on the *total* free energy, *i.e.* the sum of the two contributions. Then, the estimation of their respective weight amounts to choose a particular model for DNA. In principle, the previous method can therefore be directly transposed in our case to estimate the total potential, which is then treated in the *harmonic approximation*, *i.e.* it describes the fluctuations of the DNA bp bound to the histone core. In practice however, it is likely to be difficult, for the following reason. Within the framework described above, the potential experienced by a bp depends :

- on its position and orientation wrt the histones : if we assume that the latter are fixed, these are the "absolute" coordinates
- on the conformations of the neighbor bps, through DNA mechanics

This makes a total of 18 coordinates : the total quadratic potential is thus described by a 18x18 stiffness matrix ! Worse : this potential is also sequence-dependent through the DNA contribution, and therefore different for each trinucleotide where the considered bp is central. Let us mention that such a description implies the computation of 4752 parameters ! For comparison, we have around 30 crystal structures available for the computation of covariances, and this direct approach is therefore impossible. As a solution, one could consider to sample the phase space with MD simulations, in a very similar way as we did in Chapter 1 for naked DNA : the internal DNA contribution could then be subtracted. However, the all-atomic simulation of entire nucleosomes is in its childhood, and it is not yet clear to what extent the observed conformations are representative of the real systems, as will be discussed further later. Here, we suggest that the sequence-independent histone-DNA potential can be estimated from a limited dataset, by using the nanomechanics of nucleosomes of different sequences and conformations as a probe.

### 2.1.2 Potential of mean force

In this section, we describe our method to estimate the histone-DNA potential from a structural dataset. The physical assumption is that this potential can be described as a quadratic function of the bp coordinates : the fluctuations of the bp positions are therefore treated in the linear response regime, similarly to the analysis of naked DNA in crystal structures. For simplicity, we illustrate the method on a unidimensional system. Fig. 2.1 (left hand-side (lhs)) shows a schematic depiction of the state of equilibrium of one bp in a particular NCP structure : the considered bp experiences an attractive force in the direction of the histones, approximated by an elastic spring of stiffness $k$, directed toward an origin $a$, and an opposite force from the neighboring bp, with a stiffness $l$ and directed toward a point $b_d$ (here $d$ stands for "DNA"). As a rough image, one may consider $x$ as the radial direction of the nucleosome, following the naive model where the histones attract the DNA toward the interior, while the bending stiffness of DNA opposes this motion. This is a simplistic view, as non-radial features were found to play an important role in the nucleosome mechanics [Tolstorukov et al. 2007], and one must keep in mind that in the real dataset, the springs are 6-dimensional, with coupling terms between the different degrees of freedom. The unidimensional depiction must therefore be understood as a toy model that facilitates the understanding of the physical problem.

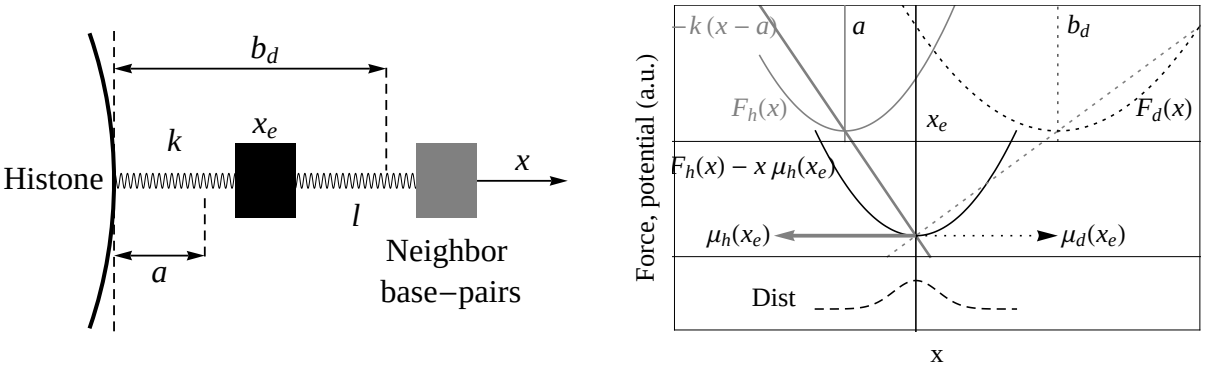


FIGURE 2.1 – Elastic model of the histone-DNA interactions. **(left)** Schematic unidimensional model : a given DNA bp experiences a harmonic opposed forces by the histone core and the neighbor DNA bp. While the parameters of the histone spring are assumed to be sequence-independent, those of the DNA spring depend on the sequence and even on the particular conformation. **(right)** Nanomechanical analysis : the sum of the histone and DNA potentials (upper panel) result in a quadratic potential centered on a conformation $x_e$, where the forces exerted by the histones and the DNA are opposed (middle panel). The thermal distribution is centered on $x_e$ (lower panel). On this figures, all quantities are shown on a single figure, with arbitrary units.

Here an important point to notice is that while the histone spring is fixed and independent of the particular DNA sequence and conformation, the DNA spring on the other hand is specific to the considered structure. If another sequence is used, the elastic parameters $l$ and $b_d$ of the base-pair step (bps) are modified. Even with the same sequence, if for some reason the neighbor bp have a different conformation, $b_d$ will be different. In the structural models, this may be the case if the NCP has crystallized in different conformations, or if a crystal contains a mutant histone or a protein bound at a remote location : the perturbation will then propagate through DNA mechanics and modify the state of equilibrium of bp bound to regular histones.

The corresponding energy landscape is described on the right hand-side (rhs) of the figure, where we use the nanomechanical framework presented in Chapter 0. The histone free energy well $F_h(x)$ (upper panel, thick gray) is a potential of mean force, and responds to an external perturbation by a internal force growing linearly with $x$ (gray thick straight line). In presence of the DNA contribution $F_d(x)$ (black dotted), the minimum of energy $x_e$ (middle panel) is such, that the two mean forces $\mu_h(x_e)$ and $\mu_d(x_e)$ are equal and opposed, which can be described geometrically as the crossing of two lines describing the mean force of either potential, of slope $k$ and $-l$ respectively. This construction assumes that the harmonic approximation is still valid for both potentials at $x_e$.

Our objective is to determine the unknown parameters of the histone potential, $a$ and $k$. A single crystallographic structure provides a value $x_{exp} \simeq x_e$. In this formalism, it is easy to see that a distribution of $x_e$ obtained from a collection of structures is of no use, since $b_d$ and $l$ are different for each of them. To estimate the stiffness of the total potential from a conformational distribution, we should sample the fluctuations of this particular state of equilibrium by MD.

Here the trick is that we know not only the position $x_e$, but also the equilibrium position of the neighbor bp. Within a given DNA elastic model, we can therefore estimate $b_d$ and $l$, and thus the characteristic mean force curve (dotted straight line), and the external force $\mu_h(x_e)$ associated to the deformation $x_e$ : this is just a unidimensional vision of the nanomechanical analysis described in Chapter 0. It is already useful when considering a single structure, but

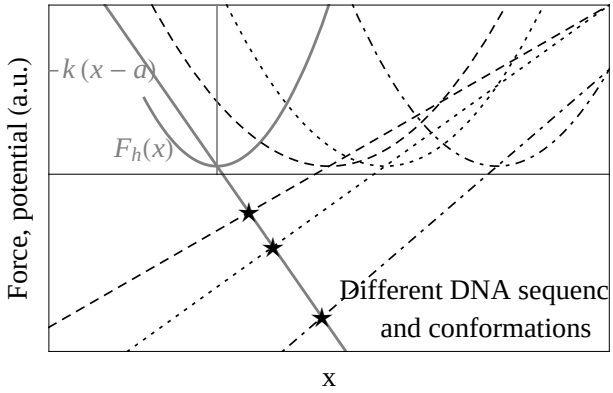


FIGURE 2.2 – Determination of the parameters of the histone-DNA elastic potential from a set of structural models. The different observed conformations (stars) result from different DNA potentials (different dashing styles), corresponding to different lines of mean force. The knowledge of the DNA elasticity provides an estimation of the DNA potential, and thus of the mean force (vertical coordinate of the stars). The elastic parameters of the histone potential (gray thick line) can then be estimated by a linear regression.

simplifies considerably the problem when several structures are compared, with different DNA wells, as shown on Fig. 2.2.

As can be seen, the points (stars) obtained from different structures (different dashing styles) align on the characteristic line of mean force of the histone potential : we can thus compute the parameters $a$ and $k$ by a simple linear regression. There is no need to assume that the distribution is representative of a "pseudo-canonical ensemble" at some undetermined effective temperature, because we do not operate a statistical analysis, but a regression. In the dataset, the bp have 6 degrees of freedom, and the histone potential is therefore modeled as a 6-dimensional harmonic spring, characterized by its symmetric 6x6 stiffness matrix $\underline{K}$ and its origin $a$, which makes a total of 27 parameters. On the other hand, each structure provides 6 datapoints, one for each dimension. Of course, it is desirable to have a large sample, with important variations in the coordinates and forces, because errors occur at every step of the procedure : uncertainties in the structural models, mapping of the atomic coordinates onto bp, elastic parameters of the DNA model and inaccuracies of the model for strong deformations...

## 2.2 Structural dataset

### 2.2.1 Structures

Since the first high-resolution mapping of the nucleosome 15 years ago [Luger et al. 1997], new structures have been obtained using the same human $\alpha$-satellite sequence or related ones, either with regular nucleosomes, or with modified or variant histones. The list of structures can be found on Table 2.1 in the Appendix. The presence of a "perturbation" (a protein or a variant histone) modifies the external potential experienced by the DNA, and the corresponding "excited" structure cannot be used to infer the nucleosome potential as described above. However, this is true only in the region where the histone conformations are modified, which is generally limited. We assumed that in the remaining part of the molecule, the analysis can be carried out. If the forces resulting from the perturbation partially propagate to remote locations

through DNA mechanics, they might even allow to sample new (and higher) parts of the energy landscape of the nucleosome. However, one must keep in mind that this could also be the case in the histones, in which case these structures might introduce erroneous datapoints.

Because all these 26 structures were obtained with essentially the same strongly positioning sequence, their ability to represent the whole conformational space of the nucleosome is questionable : they may all be in a similar specific state. Recently, additional crystallographic structures [Makde et al. 2010, Vasudevan et al. 2010] were obtained with the strongly positioning 601 sequence [Lowary and Widom 1998]. Their structural features were found to differ from the previous ones [Olson and Zhurkin 2011], and their inclusion in the dataset therefore partly solves the mentioned problem.

We have also included snapshots from Molecular Dynamics runs of the entire nucleosome based on the NCP147 sequence and structure : 5 snapshots where the thermal fluctuations give access to other excited states, and 5 "relaxed" versions of the same, obtained through short energy minimization. The snapshots were separated by 2 nanosecond ($10^{-9}$ s) (ns) in the trajectory : as a comparison, in Chapter 1, our oligomers were simulated for 50 ns. This raises the question of whether the MD-ensemble is representative of the actual nucleosomes. Possible limitations include the sampling time, which is very limited because of the huge size of the complex. Note however that this is less true than could be naively expected from the DNA length, because the tight wrapping allows to run the simulation in a much smaller box than if the oligomer was naked. Another question is whether the force fields accurately represent the DNA and histone conformations in the nucleosome. The employed DNA force fields are the result of years of calibration and improvements [Cornell et al. 1995, Cheatham et al. 1999, Pérez et al. 2007], by comparison to experimental data and quantum-mechanical calculations. Whether they can be extended to the strongly distorted conformations present in the nucleosome is an open question. This is also true for the histone-DNA conformations at the anchor points where the electrostatic effects are likely to be subtle. For all these reasons, the simple observation of a stable nucleosome trajectory is by itself satisfactory. Another problem is that while a crystallographic structure can be considered as representing the mean conformation $x_e$ of the corresponding thermal distribution (see lower panel of Fig. 2.1, rhs), this is not the case for a MD snapshot, which may be relatively far from $x_e$. As a comparison, the width of the distribution can be estimated from the B-factors seen in the data, to have a width of $\sim 5$ Å, while the error on the coordinates is estimated to be less than 0.5 Å. In the following, we treat the snapshots as if they were representative of an excited state of the nucleosome, in a metastable state of equilibrium.

For all these reasons, we propose to compare the features observed in the MD snapshots with those of the crystals at each step of the procedure. Note however that if the MD-derived structures must be taken with some circumspection, it is also the case with the structural models derived from X-ray diffraction patterns. As mentioned, a first bias comes from the sequence : the crystallization process is successful precisely because the chosen sequences induce a different behavior than standard nucleosomes, which exhibit substantial thermal fluctuations. In the crystallization process, the strong interactions between the nucleosomes of neighbor unit cells may contribute to select specific conformations. Finally, the interpretation of the diffraction patterns necessitates Molecular Mechanics models, of the same kind used in the MD simulations, and therefore subject to similar limitations if the force fields are poorly calibrated for a specific problem.

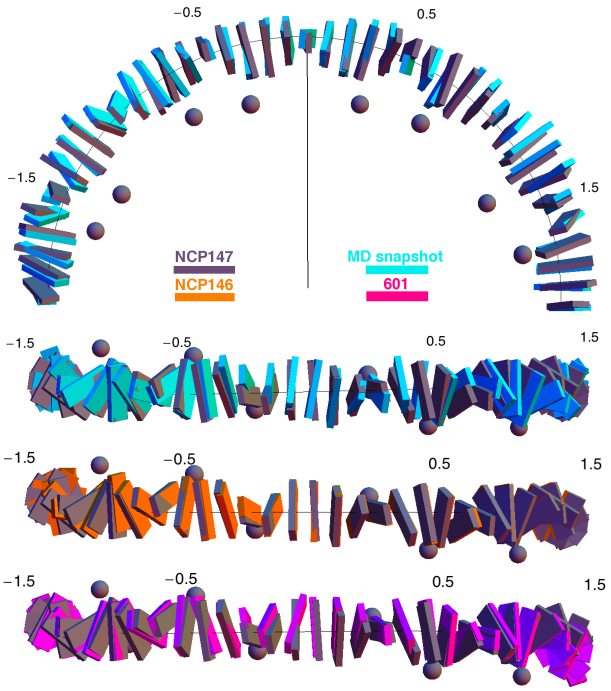


FIGURE 2.3 – Comparison of the bp conformations in the crystallographic structures and a MD snapshot. Crystal structures : NCP147 (1KX5, gray), NCP146 (1KX3, orange) and 601 (3MVD, magenta). MD snapshot (cyan) : MD1 (unrelaxed structure, see Table 2.1). The anchor points are located at the semi-integral SHL, and the primary bound phosphates in the NCP147 structure are depicted as gray spheres.

### 2.2.2 Comparison of the crystallographic structures and the MD snapshots

Before going into the computation of the forces, we compare the MD snapshots with different crystallographic structures. Fig. 2.3 shows 4 superhelical location (SHL) ($\sim$ 1/2 turn) of the NCP, for the "canonical" NCP147 structure (PDB ID 1KX5, gray), which we will consider as a reference in the remaining analysis, and for (i) one of the unrelaxed MD snapshots (MD1, cyan), (ii) the NCP146 structure (PDB 1KX3, orange), which has the same sequence as NCP147 but presents a twist defect at SHL -2.5, and (iii) a 601 crystal structure (PDB 3MVD, magenta). Overall, the bp in the MD snapshot remain remarkably close to the structure, and presents no obvious aberrant features. The deviations from the original structure are larger than in the NCP146 crystal, where they are hardly visible except in the extreme left part where the twist defect begins to appear. On the other hand, they are apparently not larger than in the 601 crystal. Whether these deviations are due to the thermal fluctuations or because the mean conformation is different than in the crystal, cannot be determined from a single snapshot. Because the important deviations are located near the anchor points, we look at these sites in more detail.

Fig. 2.4 shows the same conformations at the SHL 0.5 (left) and 1.5 (right), viewed along the superhelical axis (upper panel) and in direction of the NCP (lower panel). To emphasize the differences between the structures, the bp were reduced in size by a factor 10, which allows to add an additional relaxed MD snapshot (rMD4, green). We begin by inspecting the bp around SHL 0.5 (left), *i.e.* around the first anchor point after the dyad (black sphere at the left). The upper panel shows the displacements of the bp along the superhelical path and in the radial direction. Even at this detailed scale, the bp remain overall grouped, except at specific locations. Considering the central bp (+5), the crystal conformations are strikingly shifted outwards, in direction of the major groove, which is rather surprising from the classical view where the anchor points are the sites of *attractive* interactions with the DNA. This movement corresponds

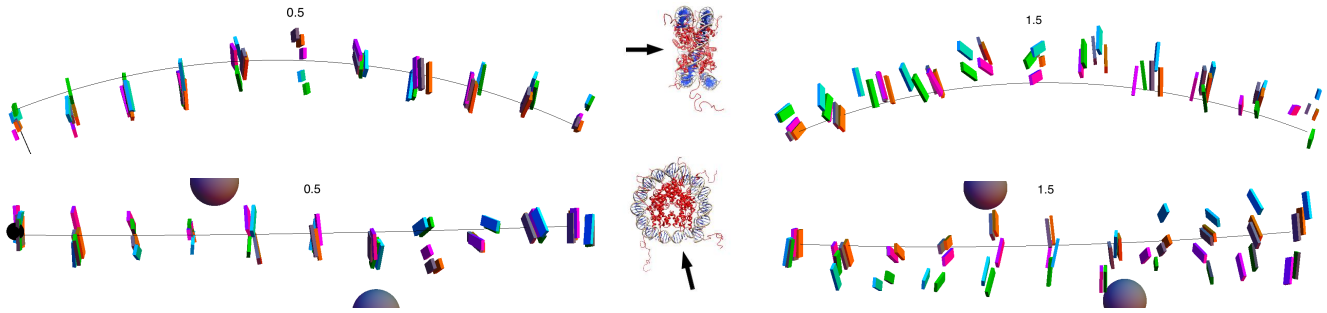


FIGURE 2.4 – Detailed bp conformations at the anchor points : SHL 0.5 (left) and 1.5 (right). Same color code as in Fig. 2.3, with the additional relaxed MD snapshot rMD4 (green). To exhibit the differences between the structures, the bp were reduced in size by a factor 10. View along the superhelical path (upper images) and in direction of the NCP center (lower images).

to the well-known peak in the shift parameter at the anchor points [Richmond and Davey 2003], and can be rationalized by considering the arginines protruding into the minor groove at this location, which could locally push the bp. This behavior is not visible in the MD snapshots, where the bp are even shifted negatively. In the lateral direction (lower panel), the crystal bp are sometimes sharply displaced, which corresponds to the slide movements responsible for the pitch of the superhelical path [Tolstorukov et al. 2007]. These movements are not always visible in the MD snapshots (two bp after 0.5), but they are also not reproducible between the structures (one bp before 0.5) [Olson and Zhurkin 2011]. At SHL 1.5 (right), the positions exhibit more deviations, both in the snapshots and crystal structures. Again, the bp are shifted outwards at the the anchor point, but this time over 3bp, and here the movement is the same in the snapshots. The lateral displacements are here sometimes more important in the snapshots (left half), but not systematically (right half). Altogether, this inspection confirms that (i) the NCP based on the 601 sequence exhibits different conformations than those with the original sequence, and thus in our framework, it allows to sample a different region of the histone-DNA interaction potential, and (ii) there is no obviously aberrant feature in the snapshots, which would justify to exclude them of our analysis.

## 2.3 Nanomechanics of DNA : computing the external forces

The nanomechanical analysis of the nucleosome has been already described in Section 2.1.2 : it allows to infer the forces responsible for a given structure. In the following paragraphs, we shortly describe the typical force patterns, which were already noticed in [Becker and Everaers 2009b], and discuss some remarkable features. Then, we compare the forces estimated on the 601 crystal and the MD snapshots.

**Visualizing the forces** The analyzed structures are models obtained from X-ray diffraction patterns, and they include a certain level of noise. When the atomistic structures are converted into rigid bp, these local irregularities result in a significant amount of noise in the force profiles. In Section 2.4, we will fit the values computed on different structures : we therefore expect the noise to average out in the fitted parameters. For the visualization of individual forces however, this important noise makes the interpretation of the profiles difficult. *For this visualization step only,* we therefore allowed the structures to slightly relax, with a maximum displacement

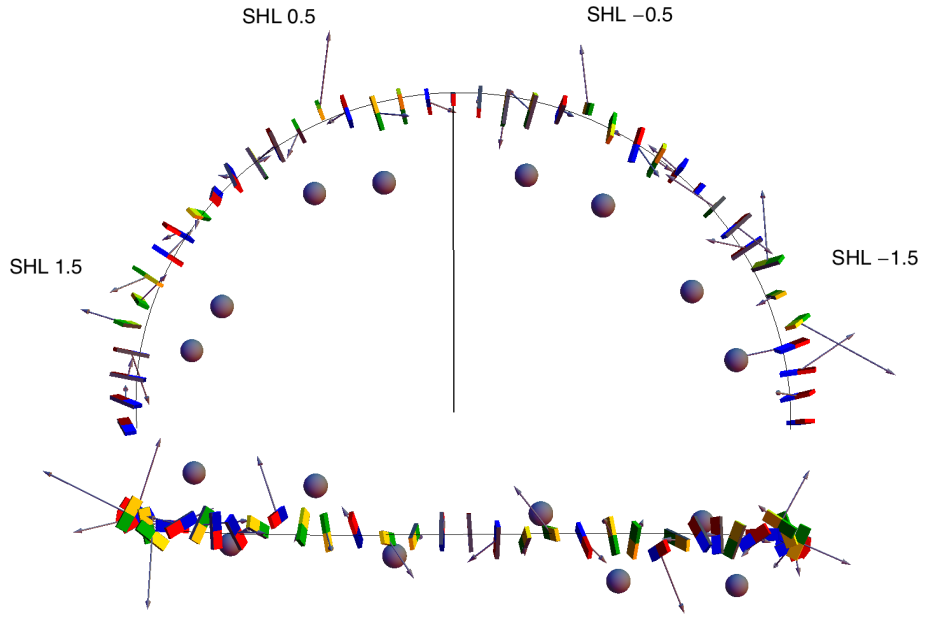


FIGURE 2.5 – Forces acting on the bp of the NCP147 structure (gray arrows) : view along the superhelical axis (upper image) and along the dyad axis (lower image). The arrow size is proportional to the force magnitude. The dyad axis is indicated by a black line, and the primary bound phosphates as gray spheres. The bp are reduced by a factor 4. The torques are not represented.

corresponding to the experimental uncertainty, following a method already used in [Becker and Everaers 2009b], and described in Appendix, section 2.6.2. The differences in the structures are hardly visible, but the relaxed profiles exhibit much more regular features, and can be compared between different structures. Unrelaxed profiles are shown in the Appendix for comparison. Note that this relaxation must not be confused with the energy minimization procedure carried on the MD snapshots at the all-atomic level. Here the relaxation is in the rbp potential, and carried on the whole structural dataset.

To illustrate the profiles, instead of looking at all force directions separately, we define a "6-norm" of the force and torque 6-vectors, which represents their total magnitude. This step implies to use a common scale for torques and forces : to do this, we define a *thermal scale* $l_{th}$, following an idea already used in [Becker and Everaers 2009b]. This scale is the ratio of the thermal torque and force scales : $l_{th} = \frac{\langle \mathbf{t}^2 \rangle^{1/2}}{\langle \mathbf{f}^2 \rangle^{1/2}} = \frac{130 pNnm}{245 pN} = 0.53 nm$. Then, for a 6-force $\mu$, by expressing the forces $\mu_{4,5,6}$ in $pN$ and the torques $\mu_{1,2,3}$ in $pN.l_{th}$, we can define the torque-force magnitude $||\mu||$ in $pN$ :

$$||\mu|| = (\frac{\mu_1^2 + \mu_2^2 + \mu_3^2}{l_{th}^2} + \mu_4^2 + \mu_5^2 + \mu_6^2)^{1/2} \tag{2.1}$$

Note that with this common scale, the torques have generally a much lower magnitude than the forces. In the 3D plots, we therefore show only the forces.

**NCP147 and NCP146 force patterns** The NCP147 and NCP146 structures are very close outside the twist defect region, and so is their force profile. The forces are shown in 3D on Fig. 2.5 for NCP147. The analysis shows that the deformations located at the anchor points are

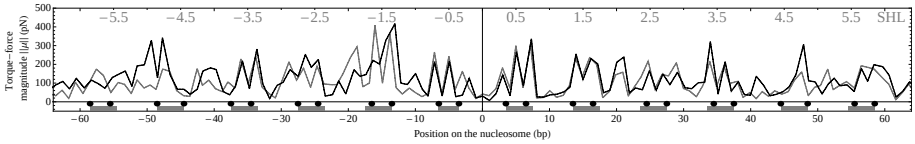


FIGURE 2.6 – Force profile of the NCP147 structure (gray) and the NCP146 structure (black) after prerelaxation. Strong forces are mostly present at the anchor points, with two different patterns : 3 successive peaks at SHL $\pm$0.5,2.5, and 2 peaks at SHL $\pm$.5. These differences may be related to the AA conformations in contact with DNA : loops and $\alpha$-helices respectively. The twist defect of NCP146 at SHL -2.5 is visible in the profile.

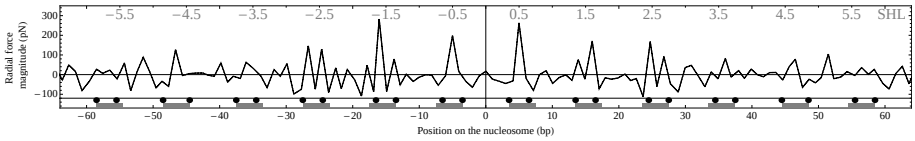


FIGURE 2.7 – Radial force profile of the NCP147 structure. A positive value indicates a force directed outwards the NCP : the anchor points are therefore also the locations of *repulsive* repulsive forces, whereas the forces are mainly attractive in intermediate regions.

indeed at least partially due to strong local forces. In the region near the dyad, the patterns are most reproducible :

- at SHL $\pm$0.5 : a strong radial force acts on the central bp, and two approximately opposed forces 2 bp away on either side, resulting in a global torque on the chain.
- at SHL $\pm$1.5 : two strong, approximately radial forces separated by 2bp

These forces can be easily distinguished on the force profile Fig. 2.6, which shows that the pattern is reproduced on other anchor points, for instance at SHL 2.5 where the molecular contacts are very similar to the ones at SHL 0.5 [Richmond and Davey 2003]. The reproducibility of the force pattern between the two different structures shows the robustness of the procedure ; the twist defect at SHL -2.5 is also visible.

Interestingly, the strong radial forces are directed *outwards* the core, which correlates with the bp being pushed away. This surprising feature is even more obvious if one plots the profile of the *radial component* of the force, Fig. 2.7 : here the anchor points appear as the points of *repulsion* by the core. Between them, the forces are slightly attractive, hence an approximately null total radial force. It is however difficult to interpret this observation, because of the complex couplings between forces and torques in DNA, which imposes to consider not only the forces, but also the torques, as well as the complex mechanical features associated with the circular shape of the nucleosomal DNA (see below).

**Force profile of the 601 structure** The detailed features of the 601 structural model differ from those of the NCP147 nucleosome [Olson and Zhurkin 2011]. This is reflected in the force profile : the forces are located in the same regions, but the patterns are generally different. Note that there are some exceptions, like the SHL $\pm$3.5, which can almost be superposed.

**Force profiles of the MD snapshots** As a test for the consistency of the nanomechanical analysis operated on the MD snapshots, one can compare the profile of an unrelaxed structure, with that of the relaxed version. Fig. 2.9(B) shows that the profiles are very similar, and the strongest peaks of the unrelaxed structure have indeed slightly decreased (with some excep-

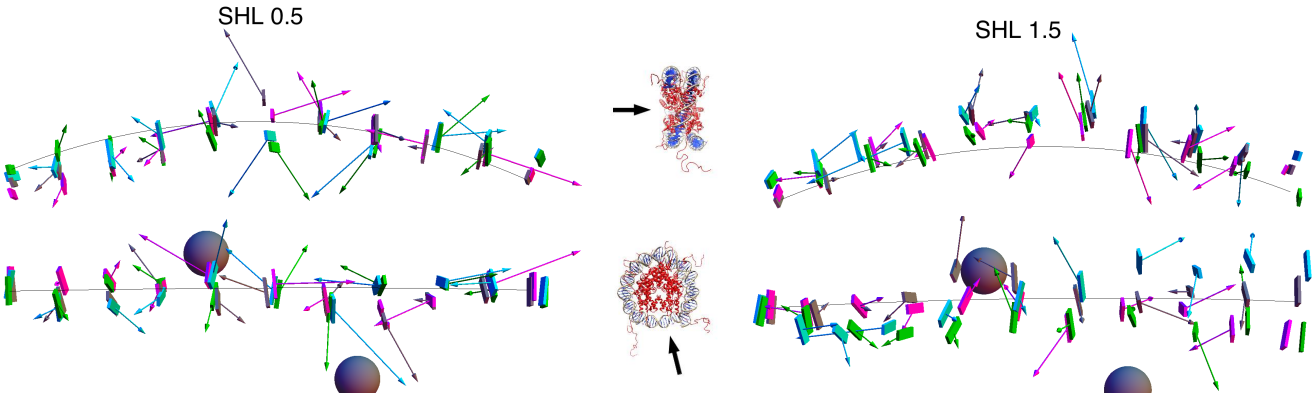


FIGURE 2.8 – Forces in the NCP147 structure (gray), the 601 structure (magenta), and two MD snapshots : MD1 (cyan) and rMD4 (green) : superhelical location 0.5 (left) and 1.5 (right). View along the superhelical axis (upper panel) and in direction of the NCP (lower panel).

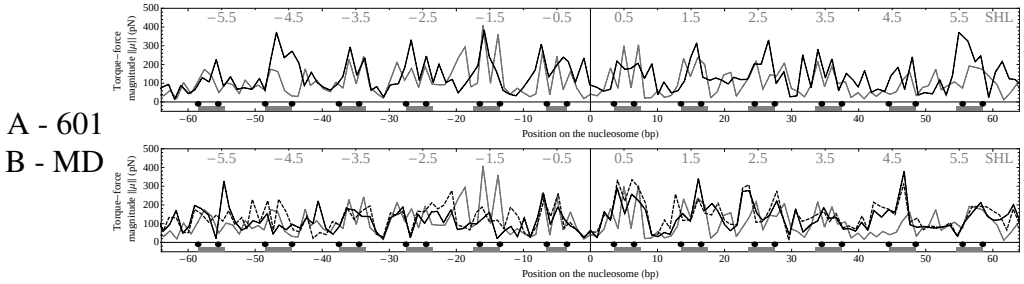


FIGURE 2.9 – **(A)** Force profile of the NCP147 structure (gray) and the 601-structure 3MVD (black). **(B)** Force profile of the NCP147 structure (gray, solid) and from the snapshot of the MD trajectory after energy minimization (rMD1, black, solid) and before (MD1, black, dashed). The forces are localized around the anchor points, as in the NCP147 structure. The force patterns are sometimes very similar (MD snapshots at SHL -0.5) but also sometimes different (601 at SHL +0.5), possibly indicating different binding mechanisms.

tions). The profiles have also many similarities with the crystal profile, as at SHL -0.5 where they can nearly be superposed. There are some exceptions however, like at SHL -1.5 where the characteristic force peaks are absent, which may be indicative of an alternative binding mode.

## 2.4 Extraction of the potential

In this section, we describe the procedure for fitting the forces obtained from the structural dataset. We remind that in the remaining, the forces are computed on the *unrelaxed* structures.

**Symmetry of the potential** The histone octamer has an axis of symmetry, the dyad axis, which passes through the center and the dyad. We therefore assume the potential experienced by the DNA to be symmetric wrt strand change.

The crystallographic structures are not symmetric however. This is straightforward for the 601 sequence which is not palindromic, and for those where variant histones make the octamer asymmetric. However, even the regular nucleosomes built with the palindromic $\alpha$-satellite sequence exhibit a symmetry breaking : for instance, the 146-bp nucleosomes have a twist de-

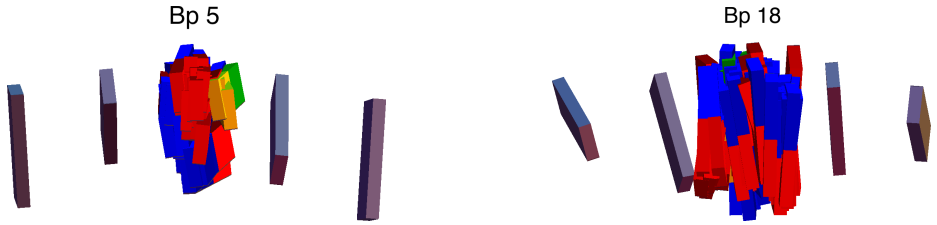


FIGURE 2.10 – Illustration of the variations in positions and orientations of the analyzed structures. The gray bp are those of the NCP147 reference structure. Colored bp : bp from the whole dataset, which are closer to the central reference bp (here invisible because buried in the group) than its neighbors. **(left)** Reference bp +5, located at the first anchor points after the dyad, where the positioning of the bp is much stronger, as well as the external forces exerted on the DNA by the histones. The differences in positions are all within the plane perpendicular to the superhelical path, and the different bp have approximately the same index on it. **(right)** Reference bp 18, located in the region between two anchor points where twist defects occur, and the positioning of the bp is weak. The scale of depiction of the bp is reduced by a factor 4. Outside the anchor points, the displacements between bp of the different structures can be important. By convention, we attribute to the same potential well, indexed by the bp $i$ of the reference structure NCP147, all bp closer to this reference bp than to its neighbors.

fect on one side, and the NCP147 structure has a slight localized asymmetry, resulting in an asymmetric force pattern.

We take advantage of these irregularities, as they give information on different conformations of the DNA under the same nucleosome potential. In other words, these alternative conformations allow to double our sampling of the potential. Thus, for each initial structural model, we consider the conformation of DNA along either strand as separate data. This makes the dataset symmetric wrt the dyad, from which a symmetric potential can be extracted.

**Grouping the base-pairs** The harmonic model for the histone-DNA interactions implies a sequence-independent elastic potential with respect to the base-pair coordinates, with each bp belonging to a different potential well (with different parameters) characteristic of the local energy landscape imposed by the histones. Thus, for an arbitrary conformation of nucleosomal DNA, where a bp may lye between the equilibrium positions of two successive bp, one must decide by convention to which well the bp belongs. Conversely, for the parametrization of the model from the dataset, this step amounts to decide which bp of the different structures belong to a given potential well and will be compared for the extraction of the parameters. The problem is illustrated on Fig. 2.10 where the bp of the different structures are shown together : (left) at the anchor points, the localization of the bp is tight along the superhelical path : the bp are all in a plane perpendicular to it (same index, see below) and thus close to the same reference bp ; (B) outside the anchor points, there are substantial variations in the positions of the bp along the superhelical path.

First, we index the bp of the different structures on a relevant scale, *i.e.* a scale describing the position of the bp with respect to the histones, in order to group together the bp experiencing the same force field. Simply counting the bp in each structure is not appropriate here, because the twist defects present in many of them generate shifts [Davey et al. 2002, Becker and Everaers 2009b]. An appropriate index is given by the position of the bp center, as projected on the superhelical path. This axis was determined from a fit of the NCP147 structure, and indexed with the dyad centered at 0, in units of the mean bps rise.

We then used the NCP147 structure as a reference, and assumed that for all the structures, the bp with an index $i-0.5 < x < i+0.5$ belong to the same potential well as the bp $i$ of the reference structure (see the groups of bp in Fig. 2.10). In other words, we assign each bp to the potential well of the closest bp in the reference structure. After this procedure, we have a single dataset for each of the 147 reference bp, which contains 0, 1 or 2 bp from each structure : 1 in the standard case, 2 if the structure is under-stretched at this location, 0 if the structure is particularly stretched and the considered reference bp is "jumped" off or if this is an excluded location because of a perturbing protein or variant histone.

This way of grouping the bp is conventional, the width of each group could be chosen differently. Some points must however be taken into consideration. The physical basis of these groups is that the bp of different structures feel the same elastic potential from the nucleosome. Obviously, this is reasonable only if they are close enough to each other. On the other hand, dividing the continuous surface of the nucleosome into discrete potential wells raises the problem of continuity : a bp lying between two reference bp will feel a discontinuous potential when passing the mid-frame.

In fact, these problems are only minor, because the 6-forces are strong and reproducible only in the anchor point regions, where the bp of all structure remain localized near the reference positions, and the problem of the boundaries between wells is irrelevant (Fig. 2.10 (A)). In fact, it has been hypothesized that the anchor points are the only regions of interactions between histones and DNA, and the intermediate DNA simply adapts to the constraints exerted remotely, without any local external force. Supporting this idea is the weakness of the forces in these regions and the lack of apparent reproducibility : the computed forces may reflect only some level of noise. This hypothesis has been used successfully to predict the location of a twist defect [Becker and Everaers 2009b].

These remarks raise the question, whether it is relevant to analyze the structures also in the intermediate regions, while we could restrict the analysis to the anchor point regions. We reverse the argument : even if the forces are weaker in these regions, and probably less crucial for nucleosome stability, there is no *a priori* argument to assume they are totally absent. In the structural models, the distance between the DNA backbone and the histone surface never exceeds 1 nm, the Debye length in physiological conditions. Thus, instead of *assuming* the absence of any interaction, we propose to *measure* this interaction (or absence of interaction) in the dataset.

**Coordinate set** For each bp $i$ of the reference structure, we now have a set $\{q_j^i, f_j^i\}$ of coordinates and 6-forces, where $1 \leq j \leq N_s$ indexes the NCP structures in our dataset ($N_s$ is the number of structural models), the coordinates $q_j^i$ are 6-vectors representing the positions and orientations, and the 6-forces $f_j^i$ are 6-vectors representing the forces and torques. The same analysis can be carried with only the positions and forces.

We first translated the coordinates and 6-forces in the local body frame of the reference base-pair. The advantage of this choice is that (i) the coordinates to be analyzed exhibit small variations around 0, rather than around a finite value following the nucleosomal geometry, thereby limiting coupling problems between rotations and translations ; (ii) the reference structure always has coordinates $\{0,0,0,0,0,0\}$, which facilitates the detection of errors ; (iii) because the positions and orientations of the reference bp are similar between different anchor points, this choice of coordinates facilitates their comparison.

Then we need to choose a coordinate chart to represent the translations and rotations bet-

ween the bp as 6-vectors. This question was shortly introduced in Chapter 0 for the rbp model, which in our implementation uses the *exponential coordinate chart*. In this system, the rotational coordinates are simply the elements of the rotation vector, but the translational part is more difficult to interpret : in particular, it depends on the rotation [Becker and Everaers 2007].

This is a problem when dealing with forces and torques : these quantities are the partial derivatives of the free energy with respect to the translational and rotational degrees of freedom respectively. In an elastic model with stiffness matrix $\underline{K}$ and equilibrium position $q_0$, it is desirable that the forces and torques experienced by a system of coordinate $q$ be given directly by the three last (respectively first) elements of the vector $\underline{K}(q-q_0)$. This is however not the case in general if the chosen system of coordinates mixes the rotational and translational degrees of freedom, as the exponential map does (for the translation).

In the particular case of the DNA mechanics under thermal motion, the molecule is rather stiff, and the deviations from equilibrium conformations remain very small for the rbp parameters : typically less than 8° for the angles. In the limit of small angles, the translational part of the exponential coordinates does coincide with the translation vector, and this problem can therefore be neglected. For the computation of the forces experienced by nucleosomal DNA, this approximation becomes already coarse, as substantial deformations can occur, especially near the anchor points [Richmond and Davey 2003].

Here, this approximation is impossible : the relative translations and rotations between the reference structure and the analyzed ones may be rather large, as illustrated in Fig. 2.10B. Therefore, we describe the bp coordinates explicitly by the rotation vector associated to their rotation matrix, and their translation vector.

**Multidimensional linear regression** From the $\{q_j^i, f_j^i\}$ of coordinates and 6-forces in the appropriate reference frame and coordinate chart, we operate a linear regression to estimate the elastic parameters of the histone core. It consists in finding the best parameters for the fit :

$$\tilde{f}_j^i \simeq \underline{K}^i(q_j^i - q_0^i) \tag{2.2}$$

where $\tilde{f}_j^i$ are the estimated forces ; that is finding the symmetric stiffness matrix $\underline{K}^i$ and the equilibrium position $q_0^i$ which minimize the quantity :

$$\sum_{j=1}^{N_s} (\tilde{f}_j^i - f_j^i)^2 \tag{2.3}$$

where $N_s$ is the number of available NCP structures. We used a global minimization algorithm implemented in Mathematica [Wolfram Research 2008] (downhill simplex method [Press et al. 2007]). For each base-pair, the algorithm computes the set of 27 parameters (21 elements of the stiffness matrix and 6 equilibrium coordinates) which minimize the distance to the dataset Eq. 2.3.

As a boundary for the domain of solutions, we constrained the equilibrium position of the potential to a certain distance away from the reference position. If the found solution was at the boundary of the domain, we repeated the minimization, imposing an iteratively increasing minimum value for the trace of the stiffness constant (doubling at each iteration, from an initially low value), until a local (and hence global within the domain) minimum was found.

As a first step, we used a very large value for the boundary distance (10 nm). In most cases, the procedure typically requires between 1 and 2 iterations, and leads to a local minimum with

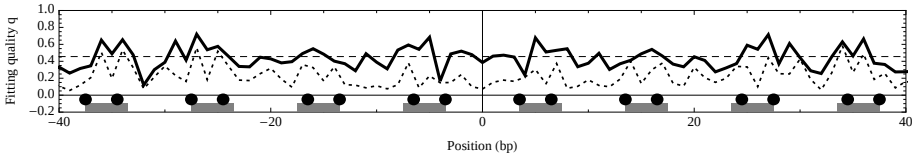


FIGURE 2.11 – Quality of the fitting model $q$, computed along the internal turn of the nucleosome (solid), as compared to a randomly generated dataset (dotted). A factor $q = 1$ indicates a perfect interpolation of the structure-derived forces, whereas a value of 0 means that the distance from the data is of the same magnitude as the force dispersion.

equilibrium position very close to the reference (between 0 and 0.5 nm) : here the effect of the constraints is very limited. For a certain number of bp however, the algorithm needs a stronger constraint on the stiffness (4 to 5 iterations) to find a minimum, which is correlated to a larger distance to the reference (in some cases, near the boundary of the domain) and a poor quality factor.

**Comparison with randomly generated data** To estimate the goodness of the multidimensional fit, we compute a generalized version of the "r-squared" correlation coefficient of a standard linear regression, obtained from the mean squared distance between the structure-derived forces and the model-derived forces, normalized by the standard deviation of the forces. Let us call this quantity $q$. There is *a priori* no lower bound to $q$ : a value of 1 indicates a perfect interpolation of the data, whereas a value of 0 means that the distance from the data is of the same magnitude as the force dispersion.

In the considered case of a multidimensional fit, involving a large number of parameters (27) on a rather limited dataset (between 20 and 30 available structures, corresponding to 240 to 360 datapoints), it is not obvious to decide what values of $q$ are indicative of a satisfactory model. To get a reference value, we generated a "control" random dataset, with the same coordinates as in the real data, and random forces taken from a normal distribution with same mean and variance as the real dataset. The comparison of the quality of both fits thus indicates directly if the measured set of forces is correlated with the bp coordinates, justifying the model.

**Results** Fig. 2.11 shows the profile of the quality factor $q$ along the internal turn of the nucleosome (solid), together with that of the randomly generated forces (dashed). We did not show the external parts of the nucleosome where the forces are less well defined, and where the values of $q$ are lower.

The fit of the real data is clearly better than that of the control : the quality factor has a mean value of 0.45, against 0.25 for the corresponding random. The fit is more precise in the anchor point regions where the bp are more localized and the forces are stronger, than in the intermediate regions where the relative noise is important. However, note that the fits on the randomly generated data also exhibit oscillations, corresponding to similar regions : this indicates that the quality factor $q$ is sensitive to the values and dispersion ranges of the coordinates and forces by itself.

Fig. 2.12 illustrates in more detail the correlations between the data and the model : here the model-derived value is plotted against the structure-derived one, dimension by dimension. In the case of a perfect agreement, all values would fall on the plotted first bisector line. The comparison with the analogous plot obtained with the randomly generated data shows again a

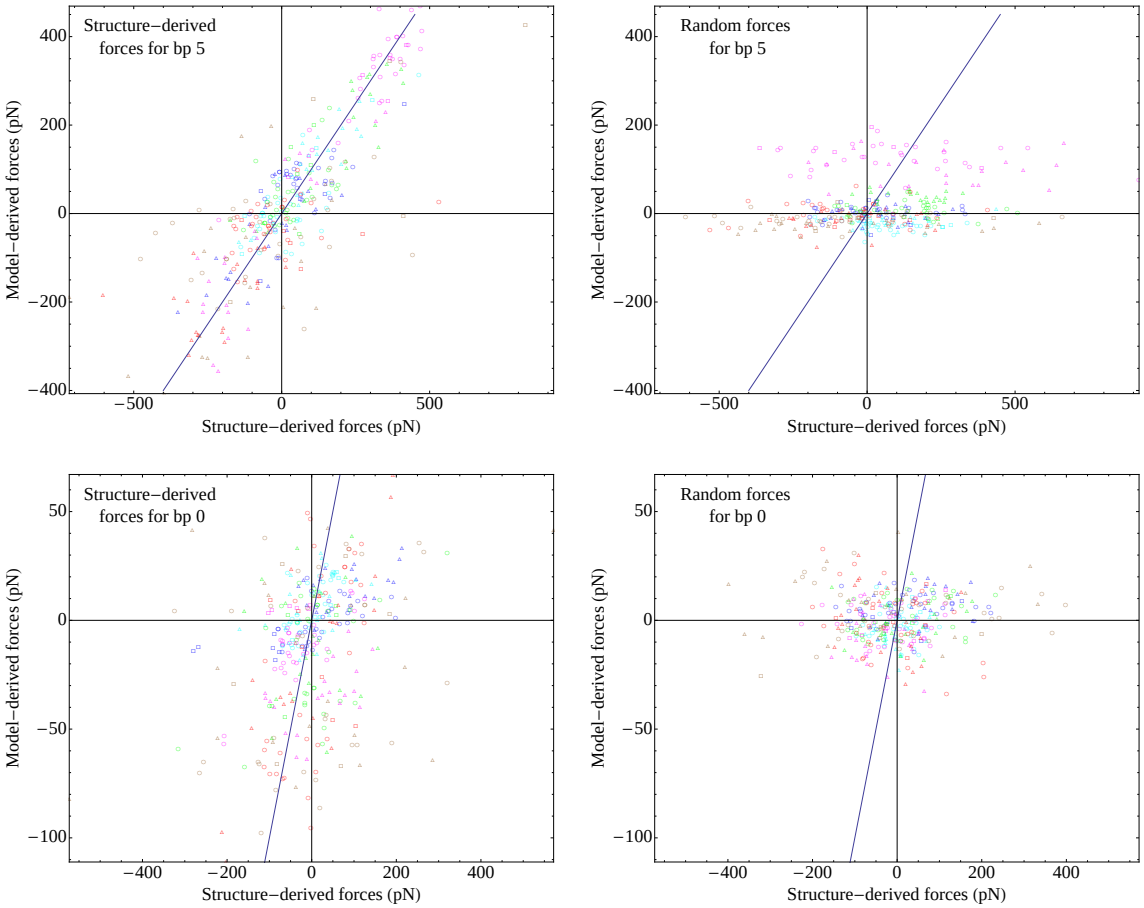


FIGURE 2.12 – Correlation between structure-derived 6-forces and model-derived 6-forces. The first bisector line is indicative of a perfect agreement. **Upper panels** : reference bp +5, located in an anchor point region where forces are strong. **Lower panels** : reference bp 0 (at the dyad), in an intermediate region. **Left** : data. **Right** : control (random). The different colors correspond to the different coordinates of the 6-forces, along the $x$, $y$ and $z$ axis in the reference body-frame : torques : blue, cyan, green ; forces : magenta, red, brown respectively. The symbols correspond to the different types of data : human $\alpha$-satellite sequence crystal (circle), 601 sequence crystal (square), MD (triangle).

better agreement, especially in the anchor point regions (bp 5).

**Repulsive forces** The very surprising result was to find that nearly all fitted matrices have negative eigenvalues : in this case, the system is not an harmonic oscillator, but an unstable quadratic potential. We checked thoroughly for sign errors, which could easily have been introduced in the many steps of the procedure. We tried to run a minimization with imposed positive diagonal values for the matrix, but the fitting was very poor. Note that the mentioned behavior does not depend on a particular reference frame for the bp coordinates, it is a physical feature invariant under frame transformation.

To better understand if the computed forces are meaningful, we consider an example where the behavior is interesting and can be checked easily. We have noticed that at the SHL 0.5, the central bp is displaced outwards the core in the NCP147 structure, and experiences a strong repulsive force. This behavior is illustrated on Fig. 2.13 (A), which is just a zoom of Fig. 2.8. Because at this SHL the minor groove faces the octamer, the displacement of the bp is a *shift* in

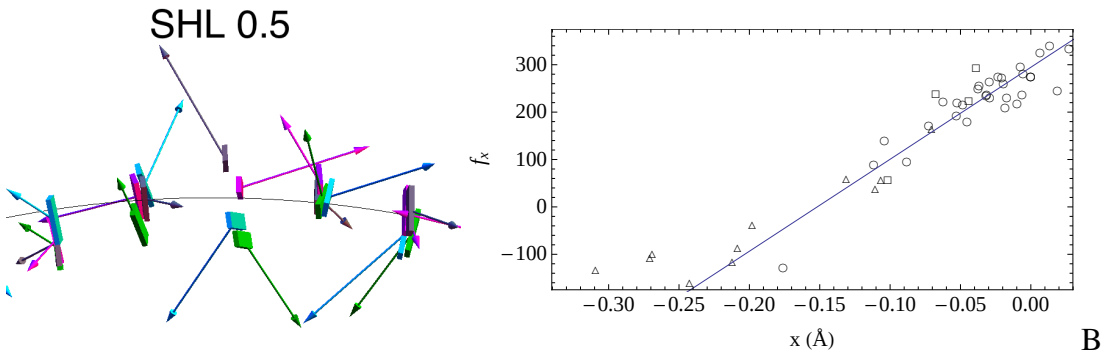


FIGURE 2.13 – Relation between the bp position and the external force, at the SHL 0.5 : example of the central bp +5. **(A)** Same colors and conventions as on Fig. 2.4 : the bp from the crystal structures (gray, magenta) are shifted outwards by strong repulsive forces, and in the MD snapshots (green, cyan), the bp are negatively shifted (inwards) and the forces are attractive. This natural relation shows that the computed external forces have no sign error. In this figure the coordinate $x$ of (B) is approximately vertical. **(B)** Coordinate-force correlation in the radial direction, for the bp+5. Symbols : $\alpha$-satellite crystal (circle), 601 crystal (square), MD snapshot (triangle). The bp divide into two groups with either strong positive radial forces (mostly crystal structures) or negative forces (mostly MD snapshots, but also a crystal). The same positive correlation is found for the whole dataset, corresponding to a repulsive quadratic potential centered at $x \simeq -0.17$.

the direction of the major groove [Richmond and Davey 2003], *i.e.* the $x$ direction in the local bp frame used in the regression. Fig. 2.13 (B) shows the estimated radial force ($f_x$ in the local reference frame) plotted against the $x$ coordinate : the majority of the points are indeed in the region around $x = 0$ in the local frame of the NCP147 bp, *i.e.* shifted toward the exterior. The 601 structures belong to this group (magenta bp on the left figure for instance), as well as most other crystal conformations. In this case, it is clear that there is no sign error in the forces, which are indeed directed toward the exterior : the DNA mechanics has some surprising features, but it is not so bizarre that when you push it, it moves back ! In this example where the forces are strong and reproducible, the slope of the fitting line is indeed the opposite of a linear spring, where it would be negative.

In a second group of structures, the bp is not shifted positively (outwards), but negatively : this group contains the majority of the MD snapshots, as can be seen in the plot (B), with two examples on (A) (green, cyan). Unsurprisingly, the corresponding forces are here negative, *i.e.* directed inwards. When discussing the structures, we suggested that this behavior could either be an artifact of the simulations, or an alternative binding mode not exhibited by the crystals. Here some remarkable features must be noted :

- not all MD snapshots are shifted negatively : some are in the group where forces are positive, although in the lower values, close to some crystallographic structures
- *a single* crystallographic structure is found in the same group of negative shifts. Together, these two observations strongly suggest that the negative shift is not a MD artifact, but indicates an alternative binding mode
- the force-displacement correlation is the same for the two groups (same slope of the fit). This means that the underlying potential is not an harmonic spring, but a repulsive quadratic potential centered around $x \simeq -0.17$, and the dataset contains equilibrium conformations on both sides of this maximum.

These findings and their consequences are discussed in the next section.

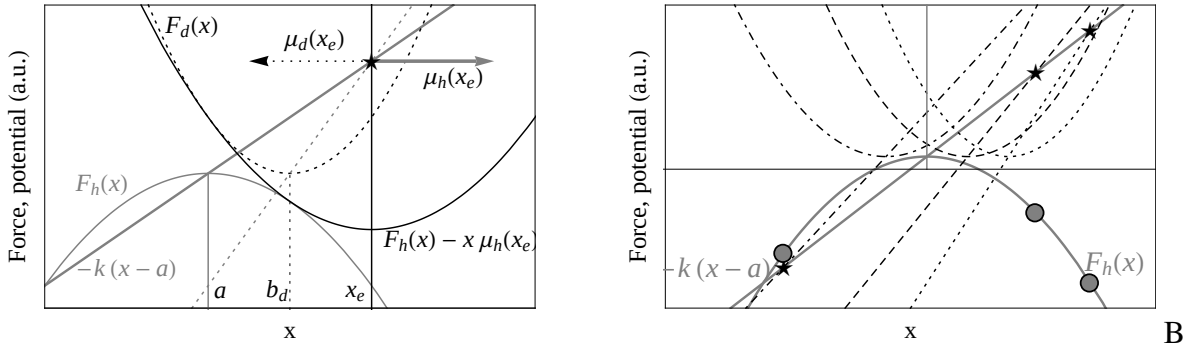


FIGURE 2.14 – Nanomechanical analysis in the case of a histone repulsive quadratic potential, and a harmonic potential for the DNA neighbor. **(A)** Single structure. The condition for equilibrium is that DNA is stiffer than the histone potential. The construction colors and conventions are identical to those used in Fig. 2.1. **(B)** Linear relation of the inferred datapoints (stars) between different structures. The corresponding points on the histone potential are represented as gray disks.

## 2.5 Discussion : repulsive potential and nucleosome stability

**Repulsive quadratic potential at the anchor points** Our analysis shows that the DNA bp experience a detectable force field from the histones. In the harmonic approximation, this potential is not a simple well, but a repulsive potential at the anchor points. We have illustrated some remarkable features in the specific case of the radial force/displacement of bp +5, confirming the absence of a sign error in the procedure. We therefore need to change the plots initially drawn to illustrate our model : the new plot looks like Fig. 2.14A. The histone potential is repulsive in the harmonic approximation, and the DNA potential is still an attractive well. The condition for an equilibrium of the total system is that the DNA stiffness be larger than the histone stiffness. Fig. 2.14B shows that with various DNA conformations, one may sample a mean force curve which has exactly the properties seen in the data.

How can we rationalize this type of surprising force field ? At least the positive part of the potential ($x > 0$) can be interpreted as the short-range repulsion of a rugged surface. If the typical radial potential of the nucleosome is similar to the shape depicted on Fig. 2.15 (A), then one may have a repulsion at the points where some Amino Acid (AA) protrude and push the DNA. The slope of the force-extension curve (Fig. 2.13) indicates that the sampled points are in the region where the curve is concave (the second derivative is negative).

**Local two-states model** This simple model accounts for the repulsive forces (positive shift), but not for the two-state distribution of position and force seen in Fig. 2.13, which is very surprising. This distribution may be rationalized, if we assume that the mechanical properties of the AA in contact with the DNA forbid the conformations where the DNA bp is approximately at rest, and must therefore "choose" between a negative and a positive shift. Fig. 2.15 (B) shows a toy model, where we would have this kind of behavior. If the connected AA are relatively stiff but allowed to freely rotate around an axis, while the DNA imposes a tight localization along the superhelical path, the state without a strong DNA deformation along $x$ is highly unstable, and we have exactly the type of potential described in the previous paragraph. In the positively shifted state (+), the bp tries to move back to the helical center-line, which is forbidden by the rigid AA : there is a strong repulsion. On the other hand, in the negatively shifted conformation (-), if the AA is not very constrained, the DNA bp feels a smaller attractive force due to the

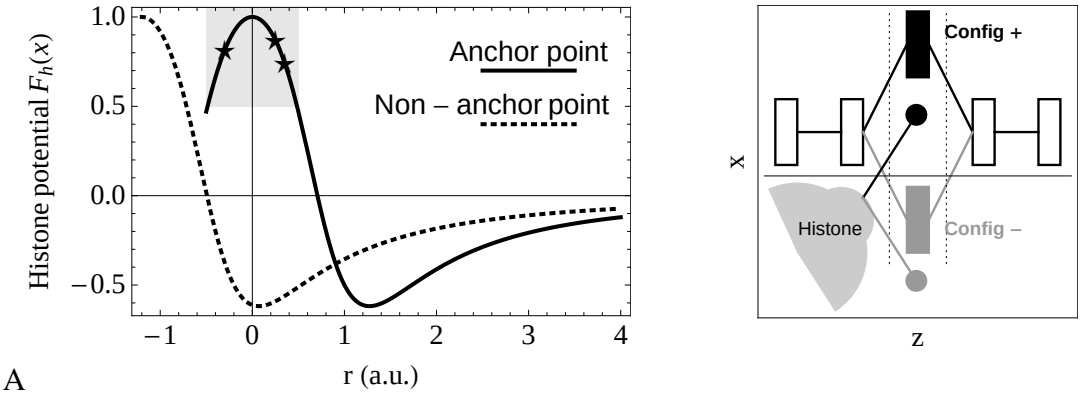


FIGURE 2.15 – **(A)** Possible schematic radial histone-DNA potential, where the datapoints sampled at bp +5 are in the repulsive shaded part, where the second derivative is negative. The potential is attractive, except at very short distance where the rugged surface is repulsive. The radius where this repulsion begins may be larger at the anchor points, where basic AA (especially arginines) are protruding. **(B)** Toy model for the 2-state distribution observed at the central bp (+5) at SHL 0.5 (Fig. 2.13 (left)). The connected AA (stick and sphere) is rigid and can only rotate around an axis. Because of the tight localization of the DNA in the $z$ direction, the bp cannot be in the helical center-line, where the AA would be compressed. It can therefore be found in two conformations : (+) the bp is shifted positively (outwards), with a strong repulsive force from the AA ; (-) the bp is negatively shifted. Assuming the constraint on the AA is here weaker, the only force exerted by the AA is the attractive force due to the positive charges, which maintain the DNA with a small negative shift.

positive charges, hence its negative shift.

**Global mechanism for nucleosome stability by DNA tension** These mechanisms give possible explanations for the observed features. How do they reconcile with the global stability of the nucleosome, if the forces at the anchor points are repulsive ? At first sight, this looks impossible. In fact, it possibly indicates a mechanism for nucleosome binding, where the elasticity of DNA has a more important role than generally considered, as we now qualitatively describe. In the most usual view, the only considered aspect of DNA elasticity is the bending stiffness, which opposes the wrapping, and is counteracted by the positive charges of the histones. In that view, the histone-DNA forces can only be attractive. A more familiar example may be helpful here : if you want to hold a stiff rubber band around a cylindrical object, you can just bend it and add some glue (positive charges) to impose sticking. Now in many such situations, we do something else : instead of bending and attaching the material, we *stretch* it : for instance, a round stretched rubber band around a bunch of parsley will not only stay in place, but even *hold* the parsley together. If you look at a small element of the rubber band, you will find that in the stretched state, the force exerted by the parsley on the rubber band is radial and repulsive ! The stability of the construction can only be rationalized at the *global* scale : it holds together because the rubber band is closed. With an open band, it can also work, if you somehow hold the ends, in order to keep the tension : in other words, it requires tangential forces. Interestingly, in some materials, the repulsive forces themselves can generate the appropriate tangential forces, taking advantage of the friction forces on a rough surface : for instance, the cellophane film used to close food containers, or similarly, plastic paraffin films used in chemistry laboratories. Here, in absence of an initial tension, the material does not stick : only when such a tension is applied by the user can the film hold on the container. The friction forces avoid the tension

to be released and make it a (meta)stable state : whether it is more or less favorable (stable or metastable state) than the fully unwrapped state is not relevant here, because the energy barrier to release the tension is important. Only when the user applies a sufficient force to remove the friction, can the tension be released and the container be opened. This is a very convenient feature if one wants to open and close the container easily and often : if you had to glue a lid to the container every time you use it, it would be painful.

This analogy illustrates a possible mechanism for nucleosome stability, where the DNA would be stretched, and the *tension* would participate in the (meta)stability of the complex. This model remains highly speculative, because we have access to *local* forces only, and the global features are difficult to estimate, but it is at least qualitatively compatible with the repulsive forces seen at the anchor points. In that view, the latter are indeed the points of DNA-histone contacts, but not necessarily the points where the attractive forces occur : rather a kind of skeleton where the DNA is held in tension. This tension could be maintained by the specific DNA-histone contacts at the anchor points, which could take advantage of the DNA structure and elasticity. The geometry of the histones approximately fits that of the DNA, with the protruding AA every $\sim$ 10bp. The interaction requires a specific orientation of the DNA helix (integral number of turns between two sites) at specific distances on the superhelical path, which imposes a constraint on the twist state of the DNA. If this state is close to, but not exactly equal to the natural twist, then there could be a twist tension, which is known to be coupled to the stretching. The specific interactions at the anchor points could thus play the same role as the surface friction in the cellophane film example, but in a much more complex way because of the 6-dimensional elastic properties of the molecule. Releasing the DNA would thus necessitate to "unbutton" the anchor points, for instance by pulling the DNA while keeping the bp in the same orientation. Of course, these hypothesis are difficult to test quantitatively, because the state of tension in the distorted conformation is difficult to estimate. Additionally, the attractive electrostatic interactions also contribute to the stability, and if the tension mechanism exists, both effects must be taken into account. Therefore, this model remains qualitative and speculative, and the next objective will be to find quantitative ways to test it on the data. Meanwhile, we just notice that it would have some satisfactory features : (i) just like like the cellophane film, it allows a rapid unwrapping of the DNA, which is released when one or several key contacts are broken, in contrast to the "glue" model where all sites of positive interactions must be unbent one-by-one ; (ii) it is easy to imagine an assisted wrapping and unwrapping mechanism, where specific proteins ("chaperones") would help to "button" and "unbutton" the anchor points ; (iii) the stretched DNA holds the histones together just like the rubber band with the parsley, which is qualitatively compatible with the observation that histone octamers dissociate in absence of wrapped DNA.

## 2.6 Appendix

### 2.6.1 List of structures

| Name | Nb | bp | Perturbation | Elim SHL | Res (Å) | Source |
|---|---|---|---|---|---|---|
| 1KX5 | 1 | 147 | | | 1.9 | [Davey et al. 2002] |
| 1AOI | 1 | 146 | | | 2.8 | [Luger et al. 1997] |
| 1KX3 | 1 | 146 | | | 2.0 | [Davey et al. 2002] |
| 1KX4 | 1 | 146 | | | 2.6 | [Davey et al. 2002] |
| 3KUY | 1 | 145 | | | 2.9 | [Davey et al. 2010] |
| 2NZD | 1 | 145 | | | 2.65 | [Ong et al. 2007] |
| 1ID3 | 1 | 146 | | | 3.1 | [White et al. 2001] |
| 2CV5 | 1 | 146 | | | 2.5 | [Tsunaka et al. 2005] |
| 2FJ7 | 1 | 147 | | | 3.2 | [Bao et al. 2006] |
| 2PYO | 1 | 147 | | | 2.45 | [Clapier et al. 2008] |
| 3LEL | 1 | 146 | | | 2.95 | [Wu et al. 2010] |
| 1P3* | 11 | 146 | H3m-H4m | $\pm$ 0.5 | 2.7 | [Muthurajan et al. 2004] |
| 1F66 | 1 | 146 | H2Av | $\pm$ 5.5 | 2.6 | [Suto et al. 2000] |
| 3AFA | 1 | 146 | H3v | $\pm$ 0.5, 2.5 | 2.7 | [Tachiwana et al. 2010] |
| 3A6N | 1 | 146 | H3v | $\pm$ 0.5, 2.5 | 2.7 | [Tachiwana et al. 2010] |
| 2F8N | 1 | 146 | H2Av | | 2.9 | [Chakravarthy and Luger 2006] |
| 3LZ0 | 1 | 145 | | | 2.5 | [Vasudevan et al. 2010] |
| 3LZ1 | 1 | 145 | | | 2.75 | [Vasudevan et al. 2010] |
| 3MVD | 1 | 145 | Protein | $\pm$6 | 2.9 | [Makde et al. 2010] |
| MD1-MD5 | 5 | 147 | | | | R. L. [private comm.] |
| rMD1-rMD5 | 5 | 147 | | | | R. L. [private comm.] |

TABLE 2.1 – Available structural dataset. The three sections are the $\alpha$-satellite-based and 601-sequence based structural models, and the MD snapshots. The dashed line separates the regular nucleosomes and the ones obtained with histone modifications or bound proteins : we indicate the particular histone and the type of perturbation : m (point modification) or v (histone variant). In the latter case, an entire helical turn of DNA was subtracted from the dataset, under the hypothesis that the histones are not strongly modified in the remaining sites : the corresponding SHL along the superhelical path is indicated in the 5th column. This is questionable for some of these modifications where the effect is important : in that case, the datapoints were only accepted when the resulting distortions were small, and eliminated when the computed forces exceeded a limiting value. We eliminated entirely 4 additional structures that were initially included, but where some bp were so distorted that the bp-mapping was impossible. The 5 MD snapshots were taken at different times of a trajectory (MD1-MD5), and a further "relaxed" snapshot was realized after 2500 timesteps of energy minimization (rMD1-rMD5). Altogether, we have 29 crystallographic structures (13 with a regular nucleosome).

### 2.6.2 Prerelaxation procedure

The analyzed structures are models obtained from X-ray diffraction patterns, and they include a certain level of noise. When the atomistic structures are converted into rigid bp, these local irregularities significantly increase the mechanical energy of the structure : for instance, the energy of the NCP147 structure (PDB 1kx5) is estimated to be $\sim 500 k_B T$, while it is only $50 k_B T$ when allowed to relax to the most favorable approximate superhelix. To reduce the noise in the extracted forces, it has been proposed [Becker and Everaers 2009b] to include an initial "prerelaxation" stage, in which the DNA bp are allowed to relax, descending the gradient of the elastic energy, under the restraint of a sharp confining potential. The size of this confinement is chosen so that the structure remains compatible with the original structural model, within the estimated atomic coordinate error. In our potential extraction procedure, we fit the forces obtained on different structures : in this case, we reason that the aggregation of data is likely to average out the noise by itself. We therefore did not implement the prerelaxation for this step. We used it only for the visualization of the forces, where the raw data are indeed painful to interpret.

This parameter was taken from the value reported in the PDB files : between 0.3 Å and 0.4 Å. Because for many structure this information was not given, we then used the value 0.4 Å. In effect, this prerelaxation selects the lower bounds of the forces and elastic energies. We found that for small ranges of confinement, the effect of this stage is mostly to scale down globally the forces and energies ; their relative values are not greatly modified by the choice of the parameter.

Here it must be noted that the precise value of the elastic energy of DNA within the nucleosome is not known from experiments : only the net free energy difference with the unbound state can be measured, which includes the histone-DNA interaction energy. As mentioned, the crystallographic structures exhibit local irregularities that considerably increase the elastic energy of the molecule ; whether these irregularities are fundamental features of nucleosomal DNA (for instance, for the strong local interactions at the anchor points) or a specificity of the few resolved crystals is not known. Thus, this value can be seen as a parameter, which defines the scale of the bending penalty : with the chosen value, the structures have an elastic energy in the order $\sim 200 k_B T$.

An unrelaxed profile is shown in Fig. 2.16.

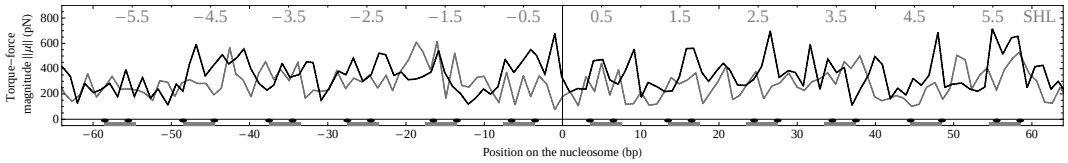


FIGURE 2.16 – Unrelaxed profiles for the NCP147 structure (gray) and the 601 structure (black) : unrelaxed version of Fig. 2.9 (A)

# Chapitre 3

# Nanoscale modeling of the nucleosomal stem

*In this chapter, we place ourselves at an intermediate level between the Nucleosome Core Particle (NCP) and the entire chromatin fiber. We combine our knowledge of the DNA elasticity at the rigid base-pair level with experimental data at two different resolutions, and we develop a model for flexible parts of the nucleosome, which are essential for the chromatin structure : the linker histone and the conformation of the linker DNA. This work was realized in close collaboration with a group of biologists, in particular Sajad Syed, Dimitar Angelov and Stefan Dimitrov [Syed et al. 2010].*

## Introduction : the linker region of the nucleosome

While detailed structural information, at almost atomic resolution, is available for the lower levels of chromatin up to the NCP, information becomes sparse at the level immediately beyond : the linker region of the nucleosome, and subsequent conformations of nucleosomal arrays.

At this larger level, thermal fluctuations prevent the use of high-resolution techniques like crystallography or Nuclear Magnetic Resonance (NMR), by displacing elements on lengthscales larger than atomic distances. This is especially a problem since these flexible elements are crucial for the interaction between nucleosomes, and in turn in the formation and maintenance of chromatin fibers and mitotic chromosomes [Makarov et al. 1984, de la Barre et al. 2001, Claudet et al. 2005] :

- the linker DNA between NCPs, comprising 10 to 90 base-pair (bp) of DNA, via their mechanical properties [Woodcock et al. 1993, Zlatanova et al. 1998, Wedemann and Langowski 2002, Schiessel 2003]
- the flexible $NH_2$-tails of the core histones, introducing subtle interactions between NCPs [Arents and Moudrianakis 1995, Mangenot et al. 2002, Dorigo et al. 2003, Muhlbacher et al. 2006, Arya and Schlick 2006, Shogren-Knaak et al. 2006, Robinson et al. 2008, Arya and Schlick 2009]
- the linker histone (H1/H5) [Fan and Roberts 2006, Zhou et al. 1998, Brown et al. 2006, Syed et al. 2010], which binds the nucleosomal DNA close to the entry/exit point of the NCP

The importance of flexible elements on fiber folding may be related to the dynamic nature

of chromatin itself : it is still under debate, whether a definite structure of the fiber exists *in vivo*. On the other hand, recent genome-wide studies of molecular factors associated to the fiber state and to a certain level of genome accessibility, such as the so-called "histone code", DNA modifications, presence of linker histone etc, exhibit a limited set of combinations of these factors [Filion et al. 2010, Kharchenko et al. 2011], which may be related to a corresponding set of fiber families. In that case, the apparent lack of regularity in the experimental characterizations of the fiber structure [Tremethick 2007] could be the result of a *static* rather than a *thermal* disorder, with a range of possible well-defined stable states. Transition from one state to another could be achieved by active chromatin remodelers.

Among the various molecular factors associated to the chromatin state, we focus here on the linker histone H1 (H1), which is a key actor in the organization of linker DNA, as illustrated by its high concentration : *in vivo*, the cell contains around one linker histone per nucleosome. It consists of a central globular domain (gH1) [Ramakrishnan et al. 1993], flanked by basic amino- and C-terminal tails, and it binds DNA close to the entry of the NCP. This results in reducing the conformational freedom of nucleosomal DNA, and bringing the two branches of the linker DNA close together, in a characteristic "stem" structure [Bednar et al. 1998]. From an electrostatic point of view, this is made possible by the positive charges of the protein (and especially the flexible tails), which may insert between the negatively charged DNAs. This aggregation may involve subtle electrostatic mechanisms (for instance, zipper-motives [Kornyshev" 1999]), as suggested by recent experiments [Fang et al. 2012] showing that upon interaction with the linker DNA, the C-terminus becomes (at least partially) structured.

The presence of H1 has been associated to dense fibers and repressed regions of the genome [Horowitz-Scherer and Woodcock 2006, Luger and Hansen 2005]. From a structural point of view, this is consistent with the hypothesis that the stem structure increases the rigidity of the nucleosomes, thereby opposing the sliding mechanism required for transcription [Schiessel 2003]. Conversely, it has been shown recently [Rochman et al. 2009] that architectural proteins (HMG), responsible for decreasing the compactness of chromatin and enhancing the accessibility of targets to regulatory factors, interact directly with the C-terminal domain of the linker histone H5 in competition with linker DNA, thereby possibly preventing the formation of the stem and subsequent compaction of the nucleosome array.

## Experimental data

Typical experimental data from our collaborators is shown in Fig. 3.1, and consists in (A) a map of the accessibility of linker DNA to hydroxyl radicals within a mononucleosome and in dinucleosomes and (B) cryo-electromicrographs (CEMs) of trinucleosomes. Both experiments have been conducted in absence of H1, and in presence of various mutants thereof (see Fig.3.2).

As established previously [Bednar et al. 1998], the micrographs illustrate the role of H1 in the stem formation (Fig. 3.1B, rows 3 and 4, additional images on Fig. 3.15) : the linker DNA paths exhibit more closed conformations, as compared to the trinucleosomes in absence of H1 (row 1). At this point, the structural study of the stem is limited by the *spatial resolution* of the experimental technique, which is of the order of the nucleosome, as that of other microscopy techniques like Atomic Force Microscopy [Montel et al. 2007]. To model the details of the stem structure, and in particular the underlying molecular mechanisms, we need information with a resolution of the order of *the base-pair*.

Such information is provided by hydroxyl radical footprinting. Hydroxyl radicals (•OH) are

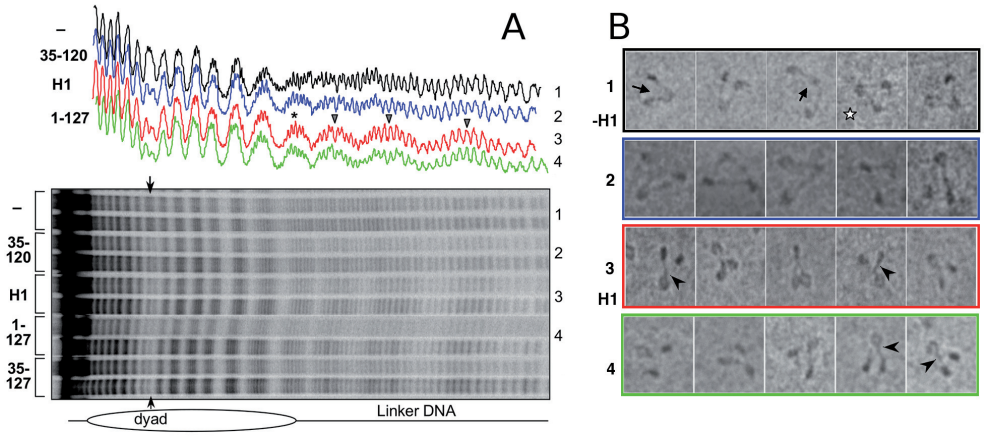


FIGURE 3.1 – Illustration of the available data (from [Syed et al. 2010]). (A) •OH-footprinting gels of mononucleosomes in the linker region, and corresponding intensity profiles : (1) without H1, (2) truncated mutant 35-120 of H1 (gH1), (3) full H1, (4) truncated mutant 1-127. The dyad region is protected by all H1 mutants, as well as the first 10 bp of the linker. Full H1 and mutant 1-127 exhibit further periodic protections on the linker. (B) Cryo-electromicrographs (CEMs) of trinucleosomes : (1) without H1, (2) gH1, (3) H1, (4) 1-127 mutant of H1. Arrowheads indicate visible stems, the star indicates a shape incompatible with the presence of a stem.

known to attack and cut the DNA backbone at a definite location (the C5' atom [Balasubramanian et al. 1998]). Here, they are put in solution with the reconstituted nucleosomes, and the resulting DNA fragments are separated by size on a gel (Fig. 3.1A). The amount of material at a given position is related to the accessibility of the corresponding site to the hydroxyl radicals within the nucleosome. Under the hypothesis that protection against cleavage is the result of the presence of an occluding macromolecule (either a histone or DNA), the "map" of protected sites along the DNA strand provides information on the three-dimensional conformation of the nucleosomal complex and the molecular contacts.

The single bp resolution was made possible by several key improvements in the experimental protocol. The precise positioning of the nucleosomes was achieved by the use of the strongly positioning "601" sequence [Lowary and Widom 1998]. Proper association of H1 with the nucleosomes was assisted by the presence of the physiologically relevant chaperone NAP-1[Shintomi et al. 2005]. Indeed, in absence of the latter, H1 binds nonspecifically to the DNA [Clark and Thomas 1986], as confirmed by the presence of a wide and concentration-dependent band on the linker DNA [Syed et al. 2010] in the corresponding gels. In contrast, in the presence of the chaperone, there is a sharp and well-defined band, at a precise location on the linker DNA, indicating a homogeneous population of nucleosomes. Importantly, increase of the concentration of NAP-1-H1 above a 1 :1 stoichiometry with respect to nucleosomes, does not change the shape of the band, which indicates that the latter are indeed all bound to a linker histone. Finally, hydroxyl radicals were used instead of other digestion agents like DNAse, which has a larger size of $\sim$10 bp, and exhibits stronger sequence-dependence.

The footprints (Fig. 3.1A) show that [Syed et al. 2010] :

- in absence of H1, the NCP exhibits a 10-bp periodic protection around cleavage, indicating the wrapping of DNA around the core histones
- binding of the globular domain of the linker histone H1 (gH1) protects the first 10 bp of the linkers as well as the DNA at the NCP dyad against cleavage.

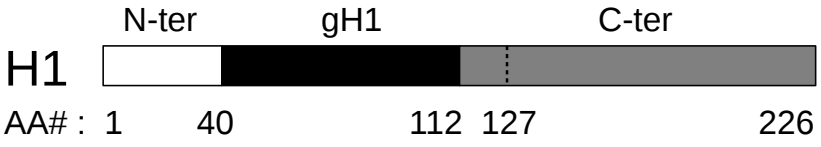


FIGURE 3.2 – Domains of the H1 histone : N-terminus (AA 1-40), globular domain (AA 41-112), C-terminus (AA 113-226)

- binding either full-length H1 or the 1-127 mutant causes in addition the appearance of a clear 10 bp repeat in the •OH cleavage pattern in the stem region of the linker DNA. Thus, the 15 first AA of the C-terminus are necessary and sufficient to induce the stem formation.

The raw experimental data clearly identify sections of the nucleosomal DNA affected by the stem formation. As in the case of scattering or NMR experiments, further interpretation of the biochemical data requires the use of macromolecular models. Our modeling is described in the following sections :

1. **Methods and models :** We give a detailed account of the analysis steps, and the parameters used in the modeling of the stem and the ensembles of mono- and polynucleosomes. *This section may be skipped for a first reading.*
2. **From footprints to a Nanoscale model of the stem :** we determine the most likely coarse-grain conformation of the H1-bound linker DNA stem in the state of maximal protection, from the protection pattern of a mononucleosome linker
3. **Soft stem structure based on nanoscale modeling of fluctuations :** we construct of a soft model for the mononucleosome stem, from a simultaneous comparison to footprinting and Cryo-Electron Micrograph (CEM) data
4. **From mononucleosome to chromatin fiber :** We compare the stem model obtained at the mononucleosome level and available data for nucleosomal arrays, in particular the footprints of dinucleosomes

Figures and large parts of the text in Sections 3.1,3.2,3.3 are taken from [Meyer et al. 2011]. The material presented in Section 3.4 is new.

# 3.1 Methods and models

This section gives a detailed account of the methods, models and parameters used in the modeling of the stem. *The reader may proceed immediately to the next sections if he wants to skip the technical details.*

## 3.1.1 Footprint analysis

The raw intensity signal (Fig. 3.1A) shows four main features :

- The smallest oscillations are single-nucleotide bands. They can be separated reliably only in a region with sufficient contrast.
- Oscillations with a period of around 10 bands reflect protection from •OH - attack.
- Trends resulting from logarithmic migration in the gel.
- The trace contains a background exposure level.

These raw traces are processed as described below. The steps of the process are illustrated on Fig. 3.6, mainly on a mononucleosome gel.

### (a) Trace alignment

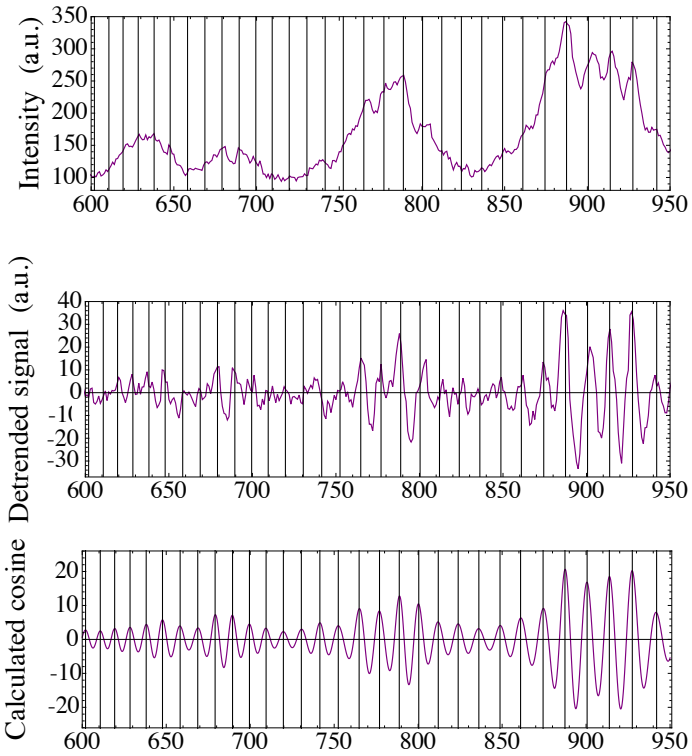


FIGURE 3.3 – Illustration of our procedure to detect the positions in pixels of the bp, on the dyad region of the NCP (30 bp). **Upper panel** : Raw data. **Middle panel** : In the first step, we subtract a moving (10 bp width) average from the original signal to isolate single bp-step variations around a more slowly varying average trend.The resulting signal is too noisy to allow for the reliable identification of individual bands. **Lower panel** : To remove artifacts we fit a sinusoidal function over several neighboring bp. Vertical bars indicate the identified band center peaks

The varying width of individual bands (7 to 14 pixels in this example) results from a combination of logarithmic migration and irregularities in the gel material. In a first step we determine the nonlinear relation between migrated distance (in pixels) and base number, as illustrated on the central region of the NCP on Fig. 3.3.

To determine the positions in pixels $x(n)$ of individual bands (numbered by $n$), we first detrended the intensity traces by subtracting a suitable moving average. We then iteratively maximized the correlation between this signal and a modulated cosine function $A(x)\cos[2\pi n(x)]$, where $A$ is slowly varying. In each iteration, the running phase of the cosine is adjusted, $n(x) \to n(x) + \delta(x)$, to improve the correlation between signal and modulated cosine, in a moving window of 7 bands length. The width of the moving window allows to assign bands even in short intermediate regions without sufficient contrast, by using the fact that band widths do not change abruptly. To ensure that no bp has been missed, we checked by eye the maximum positions and the final bp-pixel correspondence $x(n)$. The unresolved regions, where the bands could not be reliably positioned (irregular band widths), were excluded of the analysis (the resolved region of the linker is shown in the inset of Fig. 3.6, right upper panel).

### (b) Removal of gel distorsions

For the mononucleosome gels, the latter procedure was applied to each lane separately. For the higher-quality dinucleosome gels, which exhibit less distorsion, we developed an improved version where the bp-pixel correspondence $x(n)$ is computed on one lane only, and is applicable to all lanes simultaneously. This is not possible on the raw gel profiles because of local irregularities and distorsions (especially near the end of the gel), see Fig. 3.4A. We designed

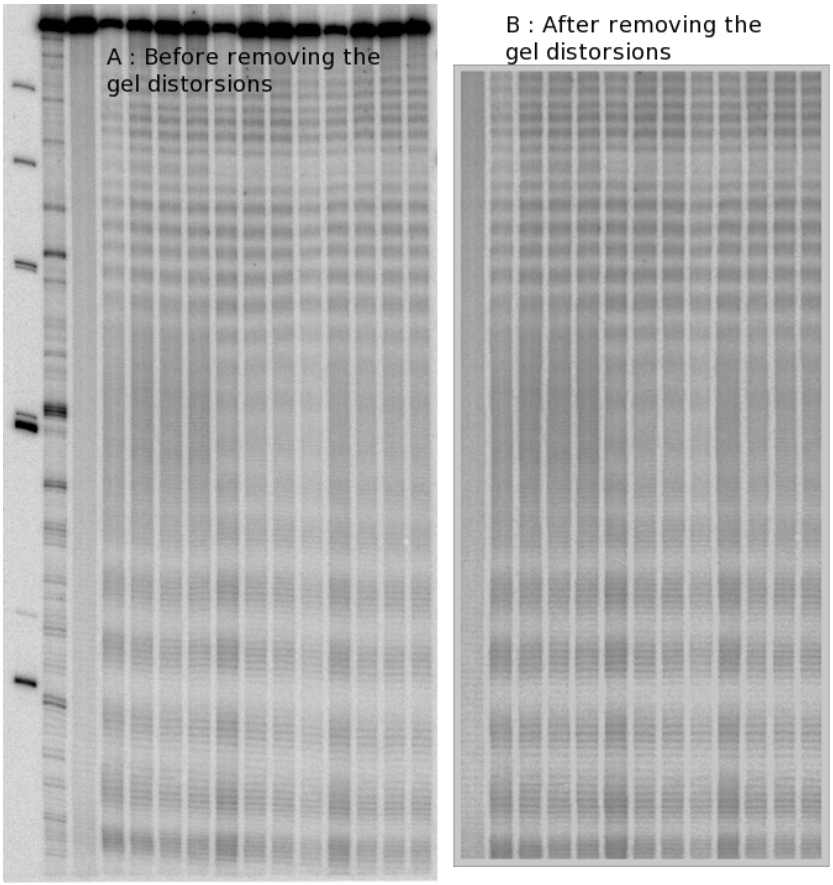


FIGURE 3.4 – Illustration of the interpolation procedure to remove the local gel distorsions on the dinucleosome gel 4, so that all lanes have the same pixel-bp correspondence. (A) Before the procedure ; (B) After the procedure, the lanes

a semi-automatic procedure to deform the image in a way that aligns the equivalent points of the different lanes, and thus approximately removes the distorsions. The procedure involves the manual placement of a grid of points located at the same position (in bp) in the different lanes, which is easy by eye. The coordinates of all image points $(x, y)$ are then transformed into new coordinates $(x', y')$, where the selected points are aligned with the reference axis : the points of the same lane have the same coordinate $x'$ and the equivalent points of different lanes have the same $y'$. All other points are transformed according to a 3rd-order polynomial interpolation of this grid. Clearly, the choice of the grid is important for the quality of the result ; in particular, there should be more gridpoints in the regions where the gel is deformed. The result of the procedure is shown on Fig. 3.4B for the dinucleosome linker : most distorsions have indeed been removed ; only in the far end region (outside the grid) does the interpolation introduce artificial features.

As a consequence, it is possible to choose on which lane we apply the bp localization procedure for the entire gel. We have considered the use of the control lane with naked DNA, where the cleavage is homogeneous and thus particularly suitable for it. However, the amount of material was smaller in this case, and the limited contrast prevented a reliable identification. We therefore used the lane with nucleosome-bound DNA at highest concentration.

### (c) Absolute positioning

To relate •OH protected areas to absolute sites on the nucleosome (with the dyad bp centered at 0), rather than band numbers only, we identified absolute lengths of DNA fragments on the

gels by using a combination of molecular weight markers present in the mononucleosome gels, laser UV irradiation and Fpg glycosylase treatment [Angelov et al. 2004], and comparison with absolute positions determined in dinucleosome gels. The unique positioning of the 601 sequence on the nucleosome then allowed to assign DNA lengths to nucleosomal sites.

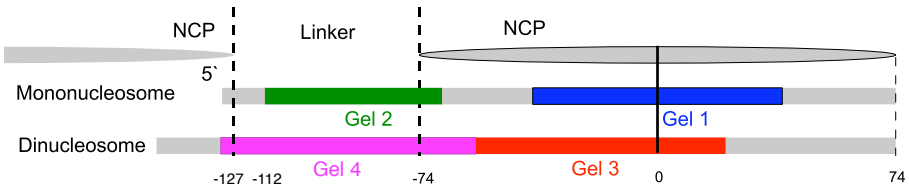


FIGURE 3.5 – Placement of the analyzed gels on the nucleosome. The gels located in the central region of the NCP (gels 1 and 3) are used to determine the absolute positioning of the identified bp (Fig. 3.6, left hand-side and Fig. 3.21). The gels on the linker region provide the information on the stem structure : gel 2 is analyzed on Fig. 3.6, right hand-side and Fig. 3.11). Gel 4 is shown on Fig. 3.4 and the processed signal on Fig. 3.16.

### Intensity per base-pair

After removing a constant level of background noise, the raw intensity signal measures the amount of DNA of a given molecular weight. By integrating over the width of each band, we obtain the irradiation intensity per band as a function of band number (since we consider only regions with well-separated bands, integration instead of fitting of multiple peaks introduces negligible errors).

### Relative accessibility

To eliminate the global trends in the trace amplitudes, we then generated a signal which represents the local accessibility of a nucleotide compared to its neighbors. In this final processing step the intensity of each bp is divided by the mean of the 3 maximum intensities in a sliding window. The window width was set to a value between 7 and 20 in the presented data. In effect, the $\sim$ 10 bp oscillations are rescaled to values roughly between 0 and 1 : see Fig. 3.6 and the final complete signal for the mononucleosome on Fig. 3.11A.

Applied to signals exhibiting no oscillations (typically, the -H1 trace on the linker), this step effectively amplifies the noise (black trace on Fig. 3.6, compare the middle and lower right panels). We therefore exclude regions where this effect prevents a reliable interpretation of the generated signal (left end of -H1 trace on Fig. 3.11A).

All processing steps were implemented in Mathematica [Wolfram Research 2008].

### Exploiting the two-fold symmetry

The mononucleosome accessibility profiles were measured for only one of the two nonequivalent strands. However, the nucleosome structure has an approximate two-fold symmetry axis, which allows to deduce an accessibility profile for the complementary strand, as follows : The 147-bp nucleosome core particle structure NCP147 [Davey et al. 2002] shows that the two-fold (dyad) axis traverses the central bp. Thus by rotating the DNA loop by a half turn around

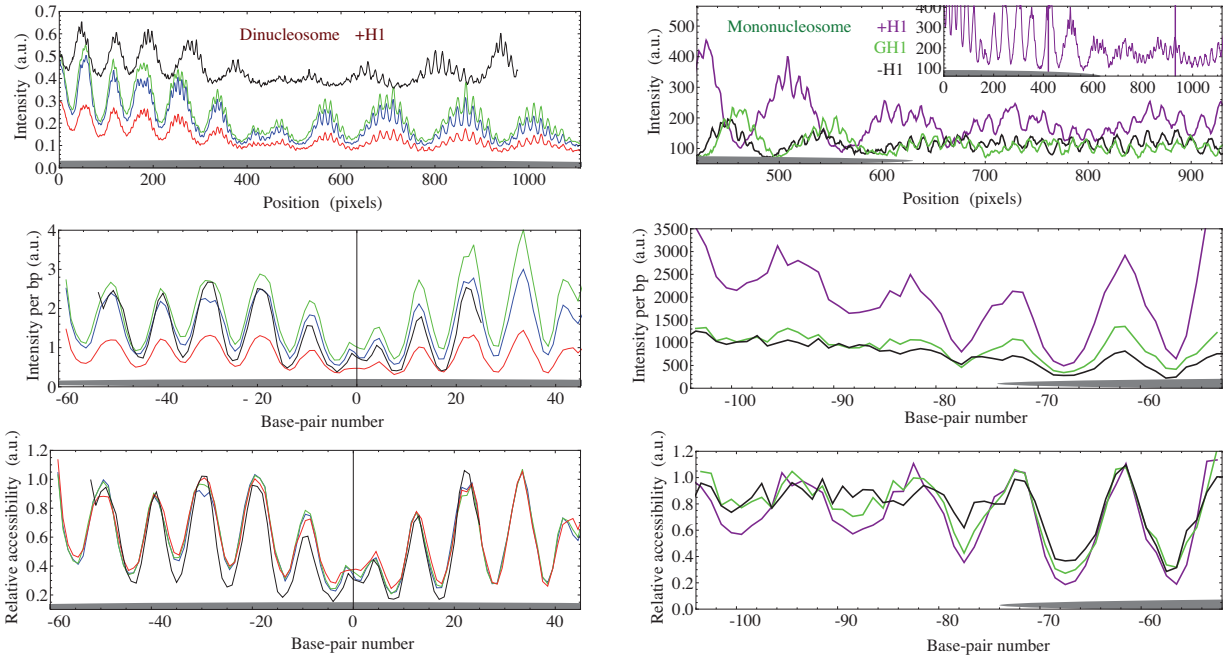


FIGURE 3.6 – **Post-processing of the footprinting gels. Upper panels** : Available raw data for di- and mononucleosomes (gels 3 and 2 respectively). Intensity variations in parts of the traces (see window in the inset, right panel) allow the resolution of bands corresponding to strands with a specific length. Outside these windows, one can only discern an oscillation with the 10bp helical period of DNA. The key step in the quantitative analysis of the gels is the identification of individual bands, *i.e.* the mapping from pixels to base-pairs (see details in the Supplementary Material). **Middle panels :** Intensity per bp, obtained by the integration of the raw signal over the bp width. **Bottom panels :** relative accessibilities : the intensity per bp is rescaled by the maxima in a moving window (of 7 to 20bp width). The resulting signal represents the relative accessibility of a site compared to its neighbors (in the same trace). **Left hand-side :** H1-bound dinucleosome traces in the NCP dyad region : red, green, blue : from the same gel, with different •OH concentrations ; black : complementary strand, from another gel. The consistency of the resulting signals (from independent traces with different relative noise) shows the robustness of the procedure (maximum difference $\sim$0.1). However, this signal does not represent the *absolute* accessibility, so that the amplitudes of different traces (or regions) cannot be directly compared (see text). **Right hand-side :** available mononucleosome traces in the linker region : -H1 (black), gH1 (green), +H1 (purple) ; the signals are shown in the resolved region where individual bands could be identified for the three signals (see window in the inset) : only one available trace for each. For a non-oscillating signal (in particular, the left end of the -H1 black trace), the last step effectively amplifies pure noise. We therefore excluded from the subsequent analysis the data for which this effect prevents any reliable interpretation. From [Meyer et al. 2011].

the dyad axis while keeping the histone core in place, one generates a second approximately symmetric conformation in which the two strands change roles.

We make the hypothesis that these two conformations are equally represented in our experiments and take the accessibility profile of the complementary strand equal to that of the measured strand (both read from 5'to 3'). This hypothesis is supported by the co-localization of protected sites from both strands (Fig. 3.11B), and by the close agreement of the two dinucleosome signals obtained independently from the complementary strands (Fig. 3.6, left hand-side and Fig. 3.21).

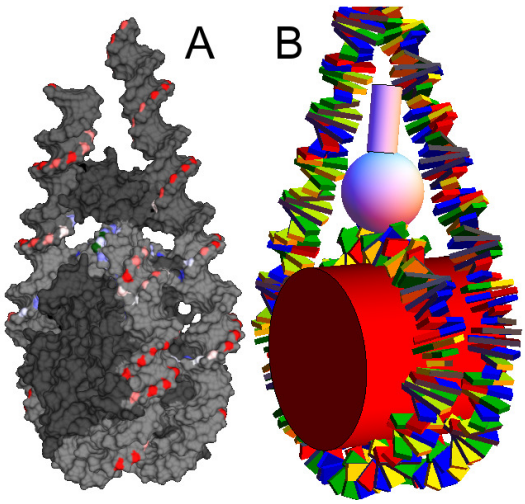


FIGURE 3.7 – Structural models used in the modeling of DNA and histones, and for the computation of structure-derived accessibility patterns. (A) Atomistic model (for the placement of gH1). NCP components from the crystal structure [Davey et al. 2002], NMR linker histone structure [Cerf et al. 1994], and straight or bent linker B-DNA pseudo-atomic coordinates [Lu and Olson 2003]. (B) Coarse-grained model (for the modeling of the stem). Rigid base-pair model (rbp) of DNA [Calladine and Drew 1997], with NCP structure obtained from (A). The histones are modeled as cylinders (core octamer and H1 tail) and spheres (gH1).

### 3.1.2 DNA and histone modeling

We have modeled DNA and histones at two different resolutions. (Fig. 3.7).

We built atomistic models where corresponding experimental information was available : on the nucleosome core and for gH1 [Cerf et al. 1994, Fan and Roberts 2006, Zhou et al. 1998, Brown et al. 2006]). For the core DNA, we used the NCP147 nucleosome structure [Davey et al. 2002]. Linker DNA was added in straight regular B-DNA conformation or in bent conformation, depending on the particular nucleosome model. These were constructed using a purpose-built Mathematica library for rigid base-pair (rbp) DNA manipulations [Calladine and Drew 1997] (with the "MP" hybrid parameter set as described in [Becker et al. 2006, Becker and Everaers 2007] from [Olson et al. 1998, Lankas et al. 2003]), with the 601 sequence used in the experiments, and then translated into pseudo-atomistic structures using the 3DNA program [Lu and Olson 2003]. We can localize the attacked sites with Ångström resolution in the molecular models. For visualization, we used software Chimera [Pettersen et al. 2004].

For the modeling of the stem, no structural models are available. We used a coarse-grained model of DNA (the rbp model, see references above). Histones are modeled as spheres and cylinders (see the paragraph "Accessibility profiles").

### 3.1.3 Atomistic models of gH1 placement

#### Three-contact model

The three-contact structure was proposed by Fan et al. [Fan and Roberts 2006] as a result of exhaustive rigid molecular docking for given DNA linker configuration. We rebuilt their structure manually by matching the orientations of the protein alpha-helices and the protein-DNA contacts for each of the three contact sites (See Fig. 3.12A). Deviating from the choice [Fan and Roberts 2006] of using a gH5 X-ray structure [Ramakrishnan et al. 1993], we considered the solution NMR structure ensemble of gH1 [Cerf et al. 1994] since it corresponds to our experimental system, has sufficient resolution for the present purpose, and allows to assess the structural variability of the protein. Specifically, while the protein fold is stable, the protein loop regions and lysine side chain orientations are highly variable ; we chose conformer 8 in the ensemble (PDB code 1ghc) since it accommodates the predicted contacts well [Fan and Roberts 2006]. At the same time, its relatively extended loop conformation does not necessitate inward bending of the DNA linkers to establish three contacts, in contrast to the somewhat more

compact gH5 conformation [Fan and Roberts 2006]. The contacting residues are Lys47, Lys51 and Ser52 (site I, orange) ; Lys63 (site II, red) ; and Lys18, Arg20, Arg72 and the C-terminal Arg75 (site III, purple). Note that these residues numbers are those of chicken H1 used in NMR studies ; they are offset by 22 AA from those in H5 used in the crystallography studies, and by 37 AA from the human H1 used in the experiments.

**Two-contact models**

Zhou et al. [Zhou et al. 1998] proposed an arrangement of linker histone onto the nucleosome based on cross-linking experiments with mutated gH5. In this model, the linker histone globular domain contacts core DNA from the major groove, at around 2 bp distance from the dyad. It also contacts one of the DNA linkers. The spatial arrangement was rebuilt by deforming one of the linkers in the DNA model, and matching the location and helix orientations of the docking solution shown in [Zhou et al. 1998] manually. We used the same molecular model (1ghc, conformer 8) for the H1 globular domain as for the three-contact model, whose shape gives close-fitting molecular contacts also in this arrangement, see Fig. 3.12B. The gH1 -helices I (cyan), II (purple) and III (magenta) are colored as in [Zhou et al. 1998] ; the C-terminal Lys75 is shown in purple, Lys63 is shown in red, contacting linker DNA. The residues Ser7,19,49 mutated in [Zhou et al. 1998] are shown in orange.

Brown et al. proposed another molecular model for linker histone placement refined by rigid docking [Brown et al. 2006]. Here the linker histone globular domain contacts core DNA from the major groove, at around 5 bp distance from the dyad, and one DNA linker. Note that globular domain positions in the two models A, B are on opposite sides of the dyad. Again the docking solution was reproduced manually by matching the reported helix orientations (different from model A) and contact residues ; it is shown in Fig. 3.12C. Residues contacting core DNA about 5 bp away from the dyad are Lys47, Lys51 and Ser52 are colored light green ; residues contacting one DNA linker are Arg20, Arg72 and Lys75 (leftmost) are colored purple. The viewing direction is the superhelical axis.

Both two-contact models should be interpreted as showing one of two symmetric coexisting configurations, forming a contact with either of the linkers.

### 3.1.4 Accessibility profiles

**Semi-quantitative •OH footprinting predictions from atomistic models**

The reactivity of attack sites is determined by their respective solvent accessible surface areas [Balasubramanian et al. 1998]. The solvent accessible surface is computed by rolling a 1.4 Å sphere representing water, or the similar-sized •OH radical, over van der Waals spheres of the atoms in the molecular model. The variations of surface accessible areas due to DNA conformation and to contacts formed with protein side-chains have been used successfully to predict the position-dependent relative accessibilities observed in •OH-footprints [Pastor et al. 2000, Strahs et al. 2003].

The corresponding MSMS program [Sanner et al. 1996] is implemented in the molecular visualization system Chimera [Pettersen et al. 2004] and can be used to compute the solvent-accessible area of each atom in a structural model (this area vanishes for interior atoms).

Lacking a thermal ensemble of structures and the resolution of single protons in our mo-

dels, we somewhat simplified the procedure, considering solvent accessible surface areas of C5' atoms directly, and using [Pastor et al. 2000, Strahs et al. 2003] 'unified van der Waals radii' [Tsai et al. 1999] to account implicitly for the hydrogens. To mimic the smoothing effect of thermal fluctuations, we increased the probe radius to 3 Å and calculated a moving average over the resulting trace with a 3 bp window. Finally, the predicted accessibility patterns for the two strands were averaged to account for the strand-exchange symmetry observed in experimental footprints.

### Semi-quantitative •OH footprinting predictions from coarse-grain models

Protection patterns for rbp models can be determined by reinserting atomic details using the 3DNA program [Lu and Olson 2003] and then following the procedure described in the preceding section. However, this approach is too time-consuming for the analysis of large ensembles of structures and we have therefore developed a procedure to estimate protection patterns on the coarse-grain level of our elastic model.

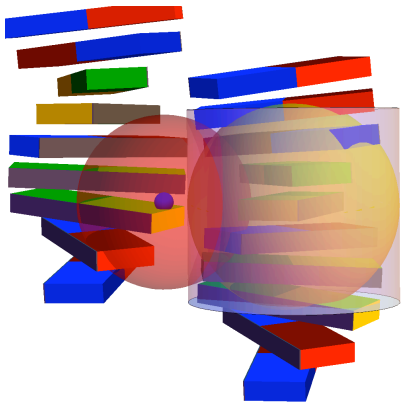

FIGURE 3.8 – Illustration of the coarse-grained structure-derived protection method. For a given C5' atom (small blue sphere), the protection value is the fraction of the pseudo-solvent-accessible sphere (red semi-transparent sphere, radius 9 Ångström) which is covered by a protecting body, here the right-hand side DNA oligomer. The protecting DNA is approximated as a cylinder (semi-transparent light pink), of radius 1.1 nm and length 2 nm, centered at the closest bp. To simplify the calculation of the overlap surface, this cylinder is approximated as a sphere (yellow sphere), tangent to the cylinder (same radius) on the segment joining the C5' atom and the central bp. In case of the protection by a histone, the latter is modeled as a sphere or a cylinder (according to the conventions described in Section 3.1.4), and the protection value is the fraction of the C5' sphere covered by this body (again, with a spherical approximation for the cylinder).

Inspired by the MSMS algorithm, we consider a "solvent-accessible sphere" around the C5' atom of each bp, of radius 9 Ångström. The protection signal associated to the bp is the fraction of this sphere covered by a protecting macromolecule, either a protein or DNA. Proteins are modeled as spheres or cylinders, and DNA as a cylinder of radius 1.1 nm, centered and oriented by the closest bp and 2 nm long. To simplify the calculation of the overlap surface, cylinders are approximated by the tangent sphere at the intersection center point and of same radius (see Fig. 3.8). Specific proteins were modeled with following parameters (Fig. 3.7 B) :

- Core histones : cylinder aligned on the superhelical axis of wrapped DNA, of radius 3.25 nm and length 6 nm.
- gH1 : Sphere centered on the dyad axis, 5.8 nm from the NCP center, and of radius 1.5 nm.
- H1 tail : Cylinder of radius 0.55 nm, and length 3 nm, tangent to gH1 sphere.

### Comparison of experiment and model-derived accessibility profiles

The simplifications in the calculation of the solvent-accessible area appear reasonable since we focus on the positions of protected sites, and aim for a semi-quantitative measure for relative (not absolute) protection. This level of precision seems adequate for a comparison to an experimental signal, which also only represents an approximation of the DNA accessibility, relative

to a free double-helix, determined via a comparison within a *local* window. It is justified *a posteriori* by the good correlation between predicted and measured accessibility profiles (Fig. 3.9). In particular, the phase of the protected sites is accurately reproduced by both methods.

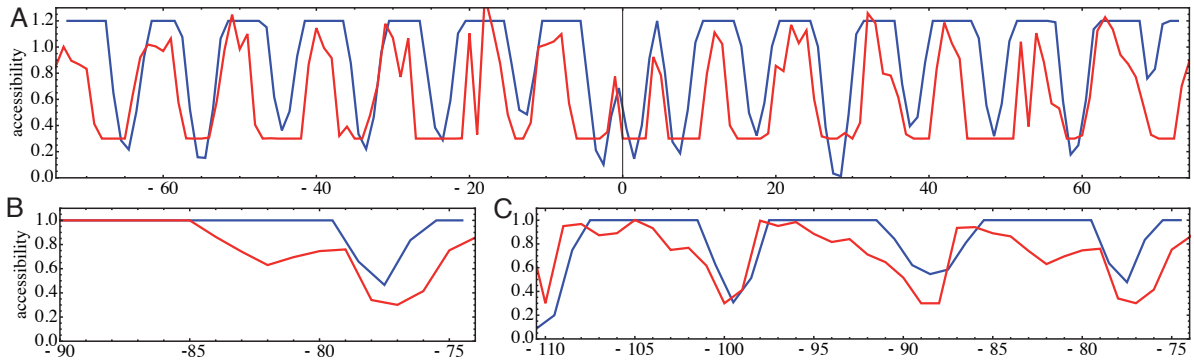


FIGURE 3.9 – Comparison of the atomistic structure-derived protection (red) and the coarse-grained structure-derived protection (blue). (A) NCP region of a H1-bound nucleosome ; (B) gH1-bound nucleosome (gH1 region) ; (C) H1-bound nucleosome (linker region). In the latter case, an arbitrary conformation of the H1 tail was chosen for the computation of the atomistic protection pattern. The protected sites are predicted at the same locations by both methods.

### 3.1.5 Stem nanomechanics

#### (a) Mononucleosome fully protected stem structure

For energy minimization, we employed the rbp model of DNA elasticity (see above). Relaxation was conducted with sequence-independent elastic parameters (a further sequence-dependent relaxation showed that the global effects of sequence-dependence elasticity were negligible in this case). DNA volume exclusion was incorporated by placing purely repulsive, truncated Lennard-Jones spheres with 2.05 nm diameter around each base pair. To enforce contacts between the two DNA linkers at corresponding maximally protected sites, linear springs were introduced between the C5' atom positions at the minima of the accessibility profile. The positions were at distances 88.5 bp, 99.5 bp and 109.5 bp on each side of the dyad (see the accessibility profile Fig. 3.11A and the corresponding springs, Fig. 3.13B).

The springs had 0.7 nm rest length (chosen to leave space for the rolling probe of the MSMS algorithm [Sanner et al. 1996]), except for the one connecting positions $\pm$88.5 bp, where the rest length was set to 1.3 nm to allow for insertion of lysines from the H1 C-terminal region. An alternative model where rest lengths were set to 0 showed only slightly different global conformations. An additional spring enforced unchanged linker separation at the height of the globular domain of H1 ($\pm$80.5 bp from the dyad), and was given a 3 nm rest length corresponding to the three-contact model (see Fig. 3.12A). The initial configuration was chosen with straight linkers (as in Fig. 3.13B), and a conjugate gradient descent was carried out until convergence, keeping only the linker-core junction bp fixed. The resulting "stem" structure is shown on Fig. 3.13C.

#### (b) Asymmetric stem model

The comparison of the experimental traces with existing models for the placement of gH1 has favored the symmetric three-contact model. We have nonetheless tested the possibility of an

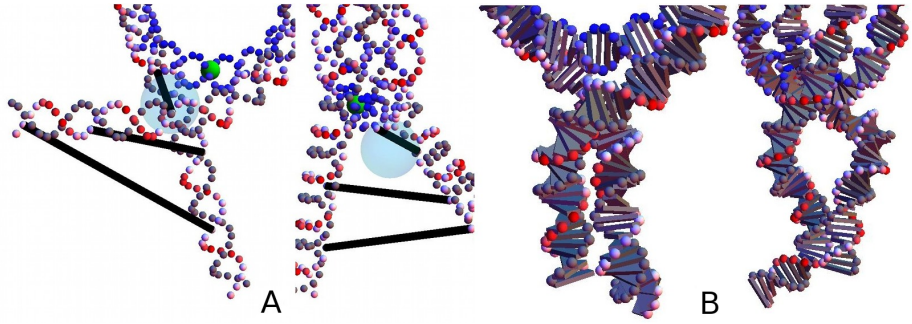


FIGURE 3.10 – Illustration of the stem nanomechanics, in the (unfavorable) case of the asymmetric, two-contact model of H1 by Brown *et al.*, which might be an alternative binding mode under external constraint. (A) The springs are placed asymmetrically, connecting the protected sites with a shift on the bent linker, consistently with the predicted placement of gH1. The first spring connects the gH1-protected site to the gH1-contact point on the NCP. (B) The resulting structure has an elastic energy of $\sim 50 k_B T$. Despite its nice "dancer" appearance, it is therefore highly unfavorable...

asymmetric stem model, corresponding to a two-contact model where only one curved linker arms contacts gH1, in the "Brown et al" conformation (Fig. 3.12), which could represent an alternative binding mode of H1 under external constraints. We reasoned that a *symmetric* distribution of springs is not compatible with an *asymmetric* stem structure. We therefore placed the first ("gH1") spring of 3 nm rest length between the protected site on the curved linker and the NCP, and two further springs of rest length 0.2 nm, connecting the 3rd and 4th protected sites of the curved linker arms with the 2nd and 3rd sites of the second linker respectively (see Fig. 3.10A). This "shift" between protected sites is compatible with an asymmetric stem. Not surprisingly, the resulting structure, shown on Fig. 3.10B, is less favorable than the symmetric stem structure based on the model by Roberts *et al.*, with an elastic energy of $\sim 50 k_B T$.

#### (c) Deformed stem models for dinucleosome modeling

For the dinucleosome, we constructed a range of deformed stem models, by imposing the contacts with some shift from the mononucleosome springs positions, either +1, 0 or -1 bp in direction of the NCP, and considering all combinations of these shifts. In this case the rest length of the springs was reduced to 0.2 nm, to ensure the contacts at the desired positions even for rather distorted structures. Accordingly, for the coarse-grain calculation of the protection on these structures, the radius of the "solvent-accessible sphere" was reduced to 0.7 nm. Comparison of the predictions for the rigid, fully-protected structures, with the experimental signal allowed to eliminate most of the structures. The only remaining ones had their contacts shifted of 1bp toward the NCP in the trunk region (see Section 3.4.1).

### 3.1.6 Fluctuating nucleosomes

In all ensembles, a part of the structure was kept rigid in the conformation obtained either from the crystal structure (-H1), with straight linkers (gH1) or from the relaxation under the experimentally obtained constraints (H1). The remaining DNA was allowed to fluctuate freely : the conformational ensemble was generated by simple Monte Carlo (see Section 0.3.2) with sequence-neutral parameters. For excluded volume calculations, DNA bp were modeled

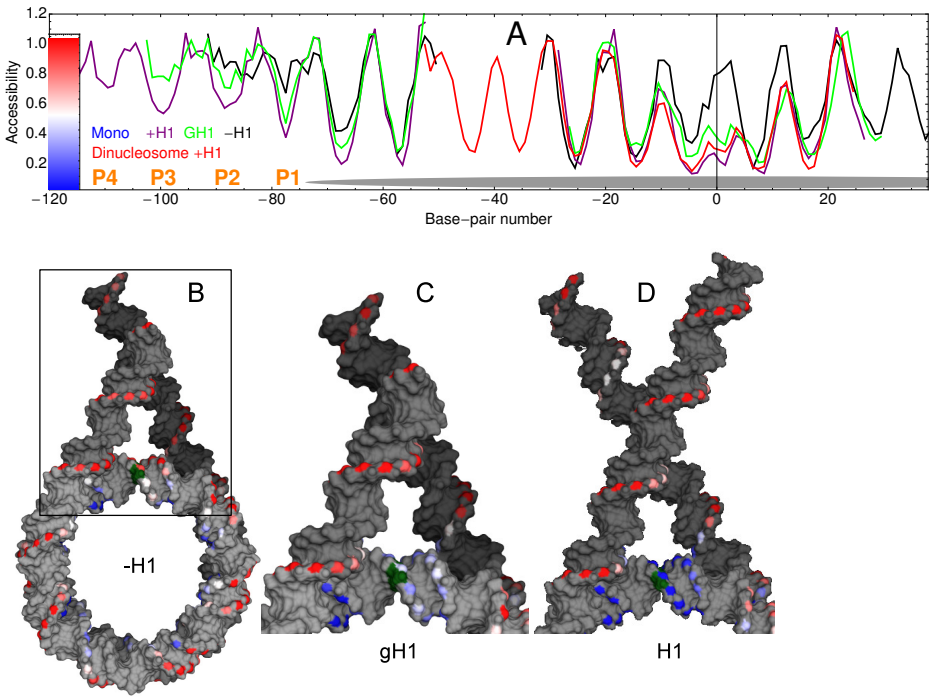


FIGURE 3.11 – (A) Relative accessibility as obtained from the processing of the gels : mononucleosome without H1 (black), with gH1 (green), with H1 (purple) and dinucleosome with H1 (red). The position of the NCP is indicated by a gray ellipse, and the protected sites P1-P4 are indicated in orange. The correspondence with the color-coding scheme used on (B)-(D) is shown on the left. On the linker DNA, the protection is weaker than on the NCP, so that protected sites appear white rather than blue. (B) Model of the nucleosomal DNA without H1 (-H1) with color-coding of the C5' atoms from blue (protected) to red (accessible). C5' atoms without footprinting data and all other DNA are shown in gray, and the dyad is indicated in green.View from the NCP superhelical axis. The protected sites are facing the histone octamer, validating the quantitative positioning of the protection trace. The co-localization of the protected sites from both strands supports the symmetry hypothesis. (C) Same for gH1-bound nucleosomal DNA, shown with straight linker DNA. New protections can be identified in the dyad region (C5' atoms facing outside the NCP), and on the first turn of the linker DNA. (D) Same for H1-bound nucleosomal DNA, with straight linker DNA arms of length 38bp, exhibiting additional protection on the linker (white spots).

as 2.2 nm-diameter cylinders : conformations where the fluctuating linkers overlap with each other were rejected from the ensemble.

**-H1** : Ensemble of nucleosome conformations without linker histone. The rigid part is the central part of the core DNA in the crystal structure (PDB ID 1kx5). To account for observed dynamical properties of the NCP, we allow a partial unwrapping of the nucleosome : the three last anchor points of each end can be detached, with a dissociation energy of $1k_BT$ per site, chosen from experimental values [Li et al. 2005, Montel et al. 2007]. The linkers fluctuate freely from the last attached anchor point. Here we also excluded configurations where either linker (modeled as a sequence of 2.2 nm-diameter cylinders (Suppl. Mat.), overlaps with the core histones (modeled as a cylinder of length 6 nm and diameter 6 nm, smaller than the most commonly accepted value of 6.5 nm). Note that the last protection of the NCP seems weaker in the -H1 case (Fig.3.1), which we interpret as a signature of partially unwrapped nucleosomes.

**gH1** : Ensemble of nucleosome conformations with the globular part of the linker histone (no tails). Following the assumption of a symmetrical binding of gH1, NCP and the first 9 bp of each linker are kept rigid (molecular contact at bp 8 from the NCP) in the conformation obtained from the relaxation, in which they contact gH1 (See Fig. 3.14B).

**H1** : Ensemble of nucleosome conformations with 1-127 mutant linker histone. The electrostatic forces responsible for the stem are localized in the H1- globular and tail region. Yet the exact extension of that constraint region is *a priori* not obvious (nor well-defined). We therefore generated three different ensembles, where the rigid region extends over different lengths in that range (16, 20 and 24 linker bp, see Fig. 3.14).

### 3.1.7 Polynucleosomes

#### Construction of polynucleosomes

**Trinucleosomes** The result of the modeling at the mononucleosome level is an ensemble of conformations, with fluctuating linkers of given length. Under the assumption that the nucleosomes of an array interact only through linker DNA mechanics and excluded DNA/histone volume, thermal ensembles of polynucleosomes can be easily reconstituted using simple Monte Carlo, at reduced computational cost, by aligning the appropriate number of fluctuating mononucleosomes. Because the model is symmetric, they can be taken in the same direction, with the linkers cut at half the desired linker length in the reconstituted array. The most time consuming operation is the test for volume exclusion between successive linkers and NCPs, where the bp were approximated by 2 nm-diameter spheres. We used this simple method for the generation of the trinucleosomes.

**Nucleosome arrays** For longer arrays of 20 nucleosomes, the same procedure can in principle be applied, however the rate of "accepted" fibers becomes extremely small when the linker length is unfavorable. We therefore implemented a naive chain growth algorithm. An iteration of the algorithm involves the random choice of a mononucleosome, which is placed on the existing chain of $n-1$ nucleosomes, and a test for overlaps with the previous ones. If the test is successful, it is integrated in the chain (now of size $n$) and the procedure goes on until the prescribed length (here 20 nucleosomes) is attained. If the test is negative, we try a new mononucleosome and the procedure is repeated $n_{trial} = 10$ times. If still rejected, we remove a nucleosome of the chain and try again. If all trials are still rejected, we stop the procedure and begin a new chain from the beginning. For each linker length, we constructed $n_{conf} = 20$ such fibers. The algorithm was not designed to accurately sample the fiber ensemble for quantitative results, but rather to generate some typical conformations for a qualitative view of the consequences of our stem modeling. In the same spirit, we computed a qualitative indicator of the steric constraint in the fiber, by computing the fraction of the $n_{conf} = 20$ started arrays which were successfully and entirely constructed.

#### Comparison of trinucleosome snapshots with images

A quantitative comparison between CEMs and model trinucleosomes is delicate, because 3D-projection as well as fluctuations contribute to the apparent variety of conformations, so

that it is difficult to avoid the introduction of a statistical bias in the selection of images.

In the CEM experiment, the 15 images of each experimental case had been selected manually from a larger ensemble of pictures, as the ones where all three nucleosomes were well-separated on the images, and the linkers were apparent. In contrast, in most images, two nucleosomes partly hide each other, or the linker DNA cannot be seen because it lies along the microscope optical axis.

To reproduce at best the experimental conditions, we first generated a large set of images of fluctuating trinucleosomes, seen from random directions. We then selected the first 15 of these images satisfying the same kind of criteria as in the experimental case.

Both experimental and model-derived images were then separated into groups according to the following criteria, ranging from "open" structures towards more "closed" ones :

- **Red** : The central nucleosome is widely unwrapped
- **Yellow** : The central nucleosome is facing, and its in- and outgoing linkers start from separate points
- **Green** : The central nucleosome is facing, the in- and outgoing linkers cross each other, and the external nucleosomes are in profile
- **Light blue** : No well-defined orientation of the nucleosomes, with seemingly short linkers
- **Dark blue** : The central nucleosome is in profile, the external nucleosomes are facing

## 3.2 From footprints to Nanoscale model of the stem

### 3.2.1 Relative accessibility and 3D structures

#### Gel analysis

The first part of our modeling is based on hydroxyl-radical footprints of mononucleosomes containing the 601 sequence ensuring well-defined positioning, and different mutants of the H1 linker histone. The proper association of the linker histone H1 (or truncated mutants) was ensured by the presence of the linker histone chaperone NAP-1, and the resulting complexes were validated by a combination of CEM and footprints. For all experimental details, please refer to [Syed et al. 2010].

The raw intensity traces of the gels are shown in Fig. 3.1A. In a first step, we extracted the relevant information (the $\sim$ 10-bp periodic protection pattern) from other features introduced by the experimental method (see Section 3.1.1). The traces were processed by automated band counting, band-wise integration and finally rescaling within a moving window. The resulting signal represents the relative •OH accessibility per nucleotide, corrected for global trends and for irregularities in the gel (Fig. 3.11A). We are able to identify the location and phasing of protected sites with single bp resolution, and to provide qualitative information on the degree of protection.

In absence of H1 (-H1), the linker DNA is fully unprotected from •OH attack. The core DNA exhibits a 10-bp (bp) periodic protection, which we attribute to the core histones. The presence of the globular part of H1 (gH1, mutant 35-112) protects the dyad region of the core DNA and a site in the first helical turn of the linker DNA (quoted site P1). Mutants with C-terminus truncated before AA 127 exhibit the same pattern. We interpret these protected sites as positions of DNA-gH1 contact or vicinity. Mutants truncated at AA 127 or further exhibit the

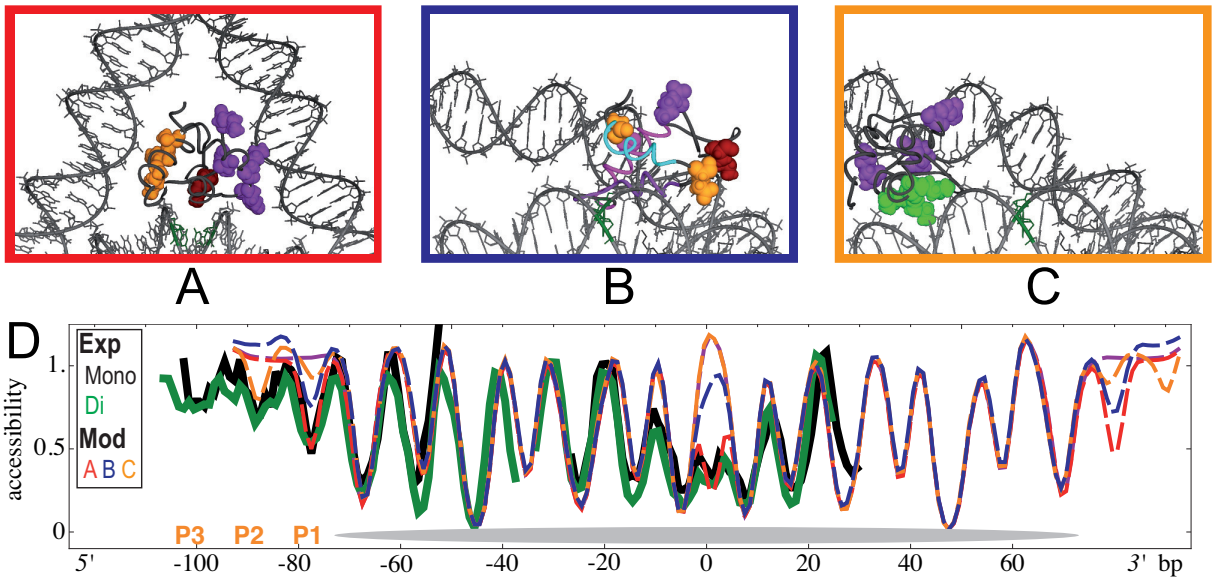


FIGURE 3.12 – Comparison of the experimentally observed protection in the presence of gH1 with relative accessibilities calculated from three different structural models for the location of the globular domain : **(A)** Three-contact nucleosome configuration by Fan *et al.* [Fan and Roberts 2006]. The viewing direction is the superhelical axis. See Section 3.1.3 for details. **(B)** Two-contact nucleosome configuration proposed by Zhou *et al.* [Zhou et al. 1998]. Contact is established with core DNA at 1-4 bp from the dyad, and with one DNA linker (the other linker is not shown). The viewing direction is the superhelical axis. **(C)** Two-contact nucleosome configuration by Brown *et al.* [Brown et al. 2006]. Contact is established about 5bp away from the dyad, and with one linker DNA. **(D)** Experimentally observed protection : thick solid lines : black (mononucleosome), green (dinucleosome). Structure-derived accessibilities : dashed lines : red (three-contact, A), blue (two-contact by Zhou *et al.* [Zhou et al. 1998], B) and orange (two-contact by Brown *et al.* [Brown et al. 2006], C), and based on a pure mononucleosome without H1 (magenta) (data already published in [Syed et al. 2010]). The predictions differ in the protection at the dyad where the two-contact models show no or very weak protection, and at the entry/exit linkers. The measured relative accessibility for a gH1-bound mononucleosome is shown in black, and that of a gH1-bound dinucleosome is shown in green.

same pattern as that of complete H1 : in addition to gH1 protections, the linker DNA exhibits a 10-bp periodic protection (P2, P3, P4). We interpret this pattern as a signature of the stem structure, in which case the linker DNA is protected either by H1 tail or by direct vicinity of the other linker branch.

### 3D-rendering of the accessibility

The •OH radicals used in the footprinting are known to primarily attack the C5' carbon atoms of the backbone sugars [Balasubramanian et al. 1998], allowing us to pinpoint protected sites in 3D molecular models of the nucleosome with Ångström resolution.

The molecular visualization package Chimera [Pettersen et al. 2004] allows the rendering of molecular structures using a color code for user-defined atom attributes. This feature was used to present the relative accessibility signals from the footprinting experiments by color coding the deoxyribose C5' atoms from blue (least accessible) over white to red (most accessible). Footprints were measured for one of the strands. Color-coding on both strands of DNA

was displayed, by exploiting the approximate two-fold symmetry of the nucleosome. Bases for which no single nucleotide resolution footprinting was available were not colored. The coloring scheme is first used to show the protection pattern of H1-less nucleosomes in Fig. 3.11B. As expected, this "3D-gel" shows directly that the 10.5-base periodicity of the experimental accessibility signal places all protected sites on one side of the double helix, facing inwards the NCP, whereas the external C5' atoms remain unprotected, as well as the linker DNA.

### 3.2.2 Molecular modeling of gH1 placement

[1] Addition of the globular domain gH1 induces additional protected sites on the core DNA and at the entering and exiting linkers (see Fig. 3.1 and Fig. 3.11 A and C. Also visible in the movies of [Syed et al. 2010], Supplementary Information). We addressed the question whether existing models of linker histone placement are compatible with the observed protection patterns. Three specific models were considered [Syed et al. 2010] : a three-contact model [Fan and Roberts 2006] where the linker histone is placed between and contacting both the entry and exit linker DNA and the dyad. Alternatively, in the two-contact models by [Zhou et al. 1998] and [Brown et al. 2006], the linker histone is placed between one linker and a site on core DNA, contacting only a single linker (Fig. 3.12 A-C).

To establish a semi-quantitative relation between these structural models and •OH footprints, we calculated footprint predictions for different structural models. C5' atoms reactivity is determined by their solvent-accessible surface area [Balasubramanian et al. 1998]. Using a simplified method inspired by existing algorithms, we were able to compare these structure-derived predictions with the measured profile, as shown in Fig. 3.12D.

The two-contact model by Brown *et al.* [Brown et al. 2006] fails to reproduce protection at the dyad. It also incorrectly predicts stronger protection at bp $-90$ than at $-80$. Since the contacts between gH1 and the core DNA at about 10 bp distance from the dyad are in the major groove, they do not protect the DNA backbone C5' atoms from •OH attack. As a result, there is no footprint of this model on core DNA at all. The two-contact model by Zhou *et al.* [Zhou et al. 1998] gives better predictions for linker DNA, generating a protected site at bp $-80$ (P1). However it fails to reproduce the strong protection pattern at the dyad, despite the proximity ; here again, protein contacts in the major groove cannot generate sufficient •OH protection. In contrast, the three-contact model is compatible with the experimentally observed protection pattern, reproducing both the characteristic double-peak dyad protection at bp 2 and the protected site at bp $-80$.

### 3.2.3 Fully-protected stem structure based on DNA nano-mechanics

For the linker DNA stem, models based on high-resolution experimental studies are not available.

We argue that under the assumption that the protected sites in the stem arise from DNA-DNA contacts [Kimball et al. 1990], knowledge of (i) the detailed register of the protected sites along the stem and (ii) DNA nanoscale structure and elasticity are necessary and sufficient to extract valuable structural information.

1. We acknowledge Nils Becker who constructed the molecular models and computed the corresponding protection patterns

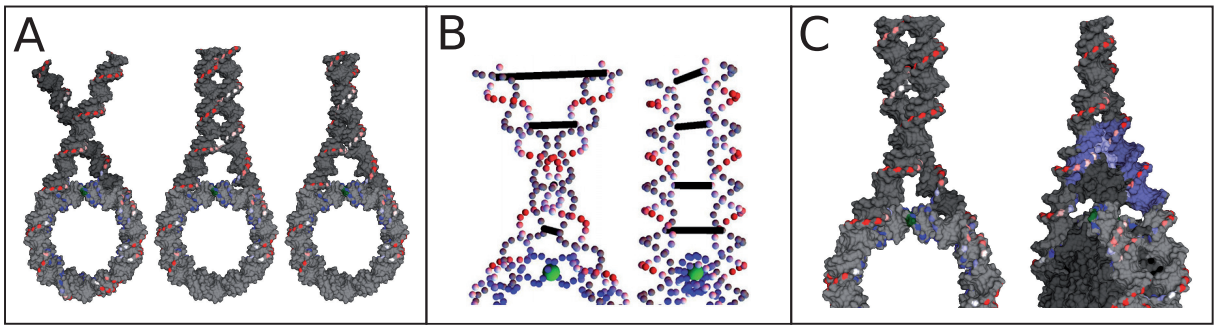


FIGURE 3.13 – Nanoscale modeling of the stem. **(A)** H1-protection color-coded nucleosomes with straight linkers (left) and two geometrical stem models. For both of them, protected sites (white spots on the linker) do not face each other. These models do not account for the observed protection pattern. **(B)** To determine the most likely stem structure compatible with the observed protection pattern, we minimized the DNA nanoscale elastic energy under the constraint that the most protected sites face each other. The initial configuration of the relaxation is illustrated here, with straight linkers. 4 springs (black cylinders) enforce contact between the protected sites. Only C5' atoms were depicted, color-coded according to the experimental protection pattern (gray when no signal). The green sphere represents the dyad. Views from the NCP superhelical axis and perpendicular. **(C)** Stem structure obtained as a result of the relaxation (already shown in [Syed et al. 2010]). Views in direction of the superhelical axis and 30° apart (dark gray histones, light blue histone tail with arbitrary conformation). The DNA is colored in blue within possible extension of the truncated H1 tail.

To see that both types of information are required, consider first stem structures with juxtaposed, weakly deformed DNA entry and exit linkers. In general, these structures will fail to create contacts at the observed sites of maximal protection, see Fig. 3.13A. Secondly, we note that any model for the center-line trajectory of the DNA linkers can be modified such that the protected sites on the two linkers face each other ; this can be achieved simply by twisting DNA appropriately along its contour. It is thus impossible to conclude on a particular shape of the linker DNA center-lines purely on the basis of the geometric arrangement of protection patterns, neglecting the physical properties of DNA.

To construct a model for the linker stem structure we therefore minimized the DNA nanoscale elastic energy under the constraint that the alignment of the linkers in space reproduces the observed protection pattern (See Section 3.1.5 for details). The resulting stem structure is shown in Fig. 3.13C. The linkers come together $\sim$20 bases outside the core particle, slightly curving into a two-start superhelical stem with a large pitch of around 100-120 bp, and extending at least to bp 40 from the NCP. This structure has, as the core particle itself, a two-fold symmetry.

## 3.3 Soft stem structure based on nanoscale modeling of fluctuations

### 3.3.1 Comparison of the fully protected stem model and the CEM images

By construction, there is perfect agreement for the location and phasing of the protected sites in the experimental traces and in the structure-derived accessibility profile of the proposed stem structure (Fig. 3.14F). There is also qualitative agreement between the calculated and the observed degree of protection, even though one might argue that the model fails to reproduce

the apparently weaker protection at the outermost P4 site. Since neither model nor experiment yield *quantitative* predictions or measurements of the DNA accessibility in the complex relative to naked DNA, it is difficult to refine the modeling or to draw more definite conclusions on the basis of the footprints alone.

However, a serious objection to the model presented thus far and in [Syed et al. 2010] can be raised from a direct visual inspection of the CEM images of trinucleosomes : stems of corresponding size can only be distinguished in some of the CEM images (Fig. 3.1B, row 3, image 2). Most images show considerable angles between the in- and outgoing linkers of the central nucleosome, *i.e.* conformations for which we would not expect any mutual protection (Fig. 3.1, row 3, images 1, 3, 4, 5 : note that the linker DNA between two successive NCPs is here 53 bp long, so that a 40 bp-long stem would fill nearly all of it. More images in Fig. 3.15).

To resolve this apparent contradiction, we note that the stem structure provided by the relaxation is the *most likely* structure accounting for protection at the observed sites. However, there is no reason to assume that the stem is rigid and always found in this conformation. Instead of considering our fully protected structure as "the" stem structure, we now develop a description of the stem in terms of an *ensemble of thermally fluctuating, partially protected structures*.

How and where do we expect fluctuations to occur ? Can we obtain meaningful results without a detailed modeling of the physical interactions responsible for the stem formation (gH1 docking, tail-linker interactions including a full exploration of the conformational freedom of the histone tails, ion mediated DNA-DNA interactions, zipper-motives etc) ?

### 3.3.2 Soft models with different ranges of rigidity

We note that the DNA deformation free energy in the fully protected stem structure is small and not uniformly distributed : 90% of a total of $2k_BT$ are localized in the first $\sim$ 20 bp of the stem, close to gH1. While the binding of gH1 alone seems ineffective, H1 variants with a small fraction of the tail induce a noticable level of stem-like associations of the incoming and outgoing linker. In modeling fluctuations in this structure, it seems reasonable to assume that the effect of H1 decreases with physical distance from the molecule.

We therefore expect the stem to open and close in a zipper-like fashion from a nucleotide around the region where H1 is localized. The simplest way to build a range of corresponding models is to divide our original, fully protected structure into two zones : completely rigid up to a variable position beyond the identified docking points with the H1 globular domain and completely free beyond. By construction, our minimal-energy structure of the fully protected stem is part of all ensembles.

We have considered three variants of the stem with 16, 20, 24 rigid bp respectively. For comparison, we also built ensembles for gH1 (9 rigid bp) and -H1 nucleosomes (allowing partial unwrapping of the core DNA, see Section 3.1.6). For each model we have generated representative mono- and trinucleosome conformations. The fluctuating part of each ensemble is the result of thermal fluctuations of DNA modeled as a rigid bp chain [Olson et al. 1998, Lankas et al. 2003, Becker et al. 2006, Becker and Everaers 2009b], with no adjustable parameter. The superpositions of aligned mononucleosome conformations shown in Fig. 3.14 illustrate the range of fluctuations in the different ensembles with a crossover from extremely floppy to very rigid linkers. For each ensemble we have analyzed protection patterns (Fig. 3.14) and trinucleosome snapshots (Fig. 3.15), to see if we can reconcile the experimental footprinting and CEM data.

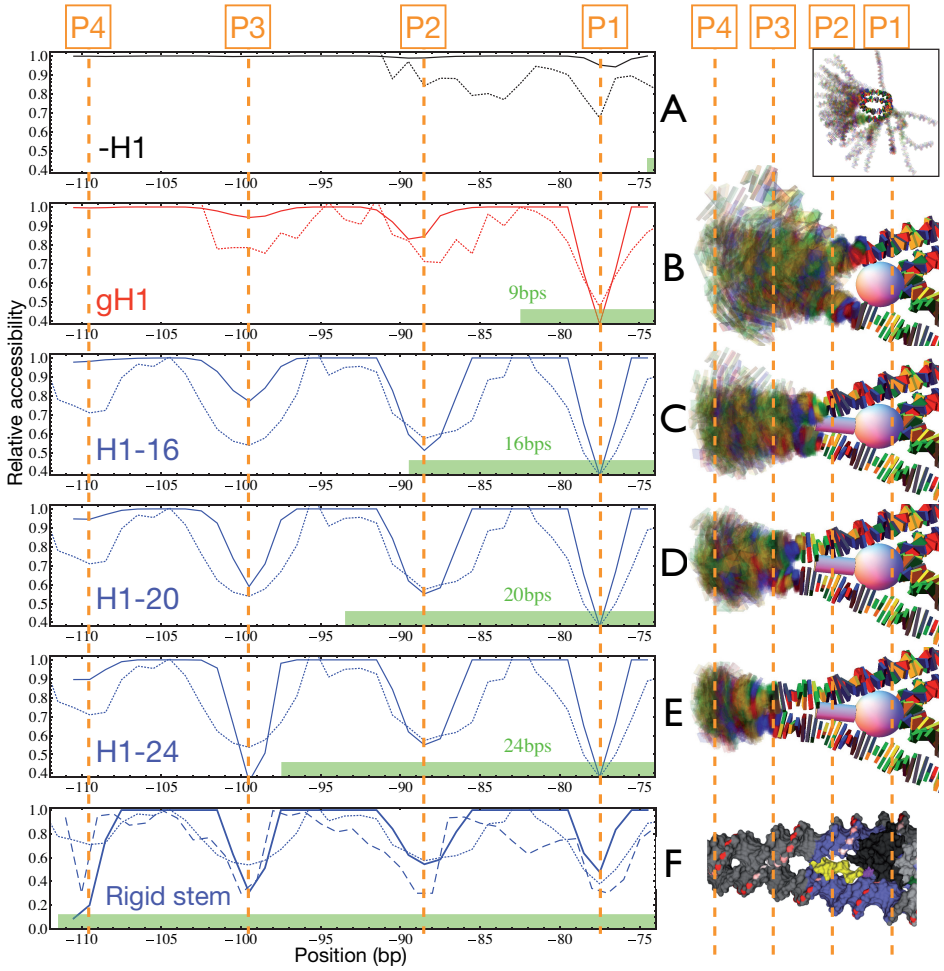


FIGURE 3.14 – Left hand-side : Comparison of the (coarse-grain) model-derived ensemble-averaged relative accessibility (solid line) with the corresponding experimental relative accessibility (dotted line). The green bar indicate the rigid part. Vertical orange dotted lines are the maximally protected sites in the experimental data. Right hand-side : Superposition of 40 snapshots of the fluctuating linkers (vertical orange dotted lines : same as on the left hand-side). (A) -H1 ensemble ; (B) gH1 ensemble ; (C)-(E) H1 ensembles with 16, 20, 24bp kept rigid respectively. As expected by construction, the most rigid stem model (24 rigid bp) reproduces the observed pattern. The apparent effect of the fluctuations is to weaken the mean protection, so that the protection of the 2 external sites (P3, P4) fades in the softest ensemble (16 rigid bp). (F) Rigid fully-protected stem structure. Left : additional dashed line : atomistic structure-derived accessibility. Right : Molecular model of the rigid stem, with DNA shown gray, color-coded protected sites (see Fig. 3.11), gH1 shown black, H1 tail shown yellow (arbitrary conformation) and DNA within possible extension of the truncated H1 tail shown blue.

The mononucleosome protection patterns shown in Fig. 3.14 were computed directly on the nanoscale structures using a coarse-grained variant of the solvent-accessible area method. As to the trinucleosome conformations, we obviously do not expect to match the cryo-EM pictures one-by-one in a comparison to a correspondingly small ensemble of appropriately projected model conformations. Rather, we have classified snapshots into five categories from open structures (red) to more closed ones (blue), allowing us to compare ensembles according to a coarse "statistics" (we give the number of configurations of each color group, red-yellow-green-light blue-dark blue). For the definitions of the groups, see Section 3.1.7. In the following we discuss the results for the various ensembles, where an increasing fraction of the linker DNA is held fixed in a rigid "root" of the nucleosomal stem.

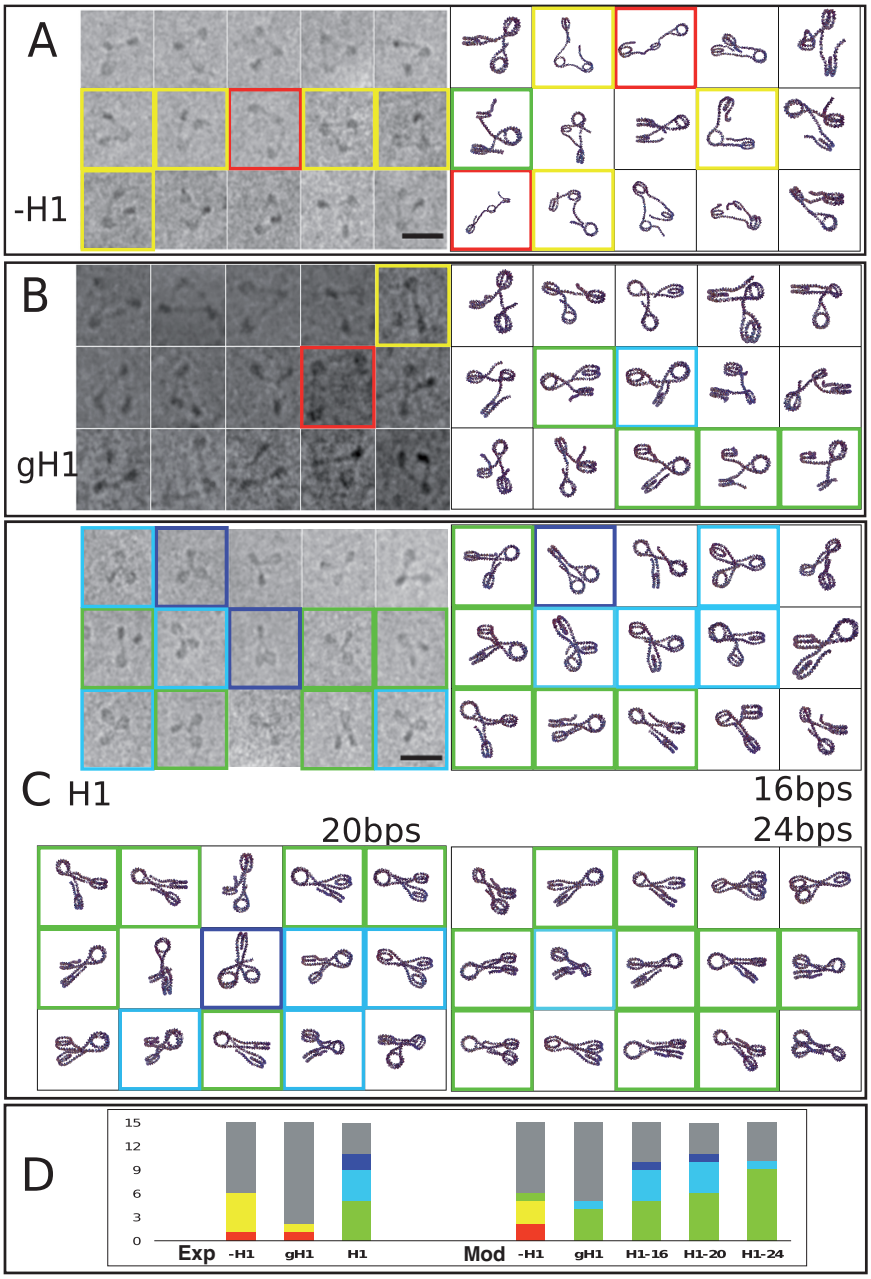


FIGURE 3.15 – Comparison of CEM trinucleosome pictures and model-derived trinucleosome snapshots, chosen randomly and appropriately projected. Images with 5 typical shapes of trinucleosomes were colored : from red (open structure) to blue (joined linkers). (A) Experimental and model trinucleosomes chosen in the -H1 ensemble. (B) Same in the gH1 ensemble. (C) Experimental and model trinucleosomes chosen in the H1 ensembles, with rigid parts of indicated lengths (smoother 16bp, 20bp, most rigid 24bp). (D) Repartition of trinucleosome shapes (indicated by colors) in the experimental C-EMs and the pictures from the model trinucleosome ensembles. Pictures which couldn't be assigned a color appear gray. Only the two softer models can account for the most open conformations observed in the pictures. These observations together with Fig. 3.14 suggest a typical rigidity extension of ∼20 linker bp, accounting for both experimental data.

**-H1**

The gels show a protected site just at the level of statistical noise, at the position of the gH1 domain (P1, -79 bp), and no other protected sites (Fig. 3.14A). The -H1 ensemble exhibits a weak protection near the same position, generated by those nucleosomes where one linker arm transiently wraps around the core histones.

Furthermore, the ensemble reproduces the characteristic open trinucleosome conformations observed in the cryo-EM experiments. The repartition of the snapshots among the "color groups", ranging from red (open structures) to blue, are : Experimental : 1-5-0-0-0, Model : 2-3-1-0-0 (Fig. 3.15).

**gH1**

Experimental traces show weakly protected sites at the same positions as the ones observed for H1 (P2, P3 in the resolved region of that gel lane). The predicted positions for transiently protected sites are those observed, generated by particular conformations in which linkers spontaneously protect each other.

The model-derived trinucleosome pictures reproduce the tendency towards more closed conformations observed in the CEM pictures. Because of the apparent variety of shapes and the poorer quality of the images, it is difficult to draw any conclusion (Exp : 1-1-0-0-0, Mod : 0-0-4-1-0)

**H1**

Fig. 3.14 C-E show the protection patterns derived from the three ensembles with 16, 20, and 24 bp constrained to their positions in our minimal energy conformation for the fully protected stem. The extension of the rigid parts are indicated by green bars at the bottom of the graphs. Within these zones, the ensembles reproduce by construction the experimental protection patterns as well as the minimal elastic energy conformation of the fully protected stem (panel 3.14F). As expected, the fluctuations reduce the protection in the outer zones. Interestingly, the transient contacts in the fluctuating ensembles are preferentially located at the experimentally observed positions. This is not an accident, but a direct consequence of the small elastic deformation energies in the outer part of our fully protected stem structure : the preferred positions of transient contacts in our model are controlled by the elastic properties of the linker DNA and the boundary conditions imposed by the rigidly held part. While the protection at P3 remains discernible in all three ensembles, the most flexible variant with only 16 fixed bp generates too little protection at P4 to be compatible with the experimental data. In contrast, the comparison of the model-derived trinucleosome images with the cryo EM pictures rather favors the softer models of the stem. The 16 and the 20 rigid bp models generate comparable distributions and varieties of structures, while the conformations from the 24- rigid bp model appear too uniform. (Exp 0-0-5-4-2 ; Mod : H1-16 0-0-5-4-1, H1- 20 0-0-6-4-1, H1-24 0-0-9-1-0).

The comparisons show that we arrive at a coherent description, if we assume that the nucleosomal stem includes 20 ± 2 bp of linker DNA : (i) closed conformations occur with sufficient probability to explain the experimentally observed protection ; (ii) the corresponding low resolution structures of tri-nucleosomes reproduce the conformational statistics of the corresponding cryo-pictures.

### 3.3.3 Discussion of the soft stem model

Based on our results, we view the nucleosomal stem as a dynamic, polymorphic, hierarchically organized structure composed of several parts :

**a root** comprising the globular part of H1 and the first 10 bp of the linkers, where gH1 preferentially establishes three contacts : one at the dyad of the histone octamer and two with the linkers, which remain on average straight and symmetric up to bp 8. The formation of the root considerably reduces the range of fluctuations of the linkers and suppresses unwrapping from the histone core.

**a trunk** or relatively rigid inner part of the stem, comprising the linkers up to bp $20 \pm 2$, in direct contact with the cationic amino-acids of the C-terminus of H1. It is in this region, that the DNA is deformed to partially align the two linkers. In our nanomechanical model, 90% of the elastic energy of the fully-protected stem structure of $\sim 2k_BT$ are located in the trunk. In our experiments, the formation of the trunk required a tail length of at least 15 AA (H1-127), indicating a biochemical control mechanism for this step.

**a flexible crown** or outer stem where the branching linkers exhibit substantial fluctuations, while preserving well-defined preferential contacts. We note that by imposing a boundary condition on the linker conformation, the influence of the trunk structure may extend beyond the region of direct interactions between H1 and the linker DNA. In particular, the linkers might appear connected without there being strong direct interactions. In our experiments, this influence extends to at least 40 bp away from the NCP, reaching the typical linker lengths of native fibers (40 bp for a *half-linker*). Our experiments [Syed et al. 2010] also suggest that the full C-terminus, while possibly stabilizing further the stem, does not qualitatively modify its structure. As a natural explanation for this effect, we suggest that the terminus may remain confined in the trunk region.

The hierarchical organization implies that the stem opens and closes in a zipper-like fashion. Under our experimental conditions, thermal fluctuations mainly affect the branching linkers in the crown. The response to external forces depends on their magnitude. While weak forces should essentially deform the crown, larger forces should disrupt the trunk and eventually the root before unwrapping the core particle [Kulic and Schiessel 2004]. Such forces may be exerted in a controlled manner in single-molecule experiments [Kruithof et al. 2009]. They also arise during the condensation of the chromatin fiber, where H1 might accommodate a fairly large range of linker conformations, if other interactions [Mangenot et al. 2002, Muhlbacher et al. 2006, Arya and Schlick 2006] compensate for the free energy cost of a stem deformation or a partial stem disruption. We note, that this view of the stem is a natural extension of the current, dynamic picture of the nucleosome core particle : instead of a passive and rigid wrapping of 147 bp of DNA in a conformation resembling crystal structure(s) [Davey et al. 2002, Makde et al. 2010], DNA and histone octamers form a highly dynamic complex, where DNA spontaneously unwraps and rewraps from the ends [Li et al. 2005, Koopmans et al. 2007, Montel et al. 2007] with actively created [Shukla et al. 2010], diffusing [Kulic and Schiessel 2003b,a] defects ensuring mobility [Schiessel et al. 2001].

How do our results match those of previous studies ? There is a remarkable diversity of experimental [Zhou et al. 1998, Brown et al. 2006] and modeling [Bharath et al. 2003, Fan and Roberts 2006, Wong et al. 2007, Cui and Zhurkin 2009, Pachov et al. 2011] results for the mode in which gH1 binds in what we now refer to as the "root" of the stem. The structures presented here are derived from experiments where mono-, di-, and tri-nucleosomes were reconstituted following a carefully elaborated protocol recreating *in vivo* conditions such as the presence of the chaperone NAP-1. If we believe them to represent a free energy minimum for systems dominated by intra-stem interactions, it is an interesting question, whether other binding modes might serve to stabilize alternate structures in nucleosomes under external constraints. If one interprets the multitude of predicted gH1 binding modes in the root as a *feature* of the molecule and not as a failure of the employed modeling schemes, then the ability of the stem to adapt to external constraints might be even larger than apparent from our experiments.

Little was known about the trunk region. Neglecting fluctuations, Bharath *et al.* [Bharath et al. 2003] used structural analogies and bioinformatics methods to predict a placement of gH1

close to the 2-contact model of [Zhou et al. 1998]. For the trunk, they proposed a particular conformation of the C-terminus making contact with the linker DNA up to 24 bp away from the NCP, beyond which the sharply bent linkers were supposed to diverge in the crown. While the predicted extension of the H1-DNA contacts is rather close to that of our trunk, the details of the proposed stem structure are incompatible with our experimental results. This hold for the resulting protection pattern close to the DNA dyad as well as for the predicted divergence of the two linkers beyond the contact zone with the H1 tail, which is difficult to reconcile with the experimentally observed protections P3 and P4 (see Fig. 3.14).

## 3.4 From mononucleosome to chromatin fiber

As discussed above, the determination of the stem structure is particularly relevant with respect to the putative structure of the chromatin fiber. The path of the linker DNA in the latter is still unknown today, both *in vivo* and *in vitro*, and is subject to a longstanding debate [Schiessel 2003]. Only for a model tetranucleosome [Schalch et al. 2005] could this path be resolved, in a very specific case where the steric hindrance between nucleosomes and the distorsion energy of the linkers (without linker histone) select a relatively rigid conformation allowing for X-ray crystallography, and consistent with the crossed-linker (or "zig-zag") model of the fiber.

The physical interactions involved in the chromatin fiber folding can be divided into two categories :

- *intranucleosome interactions*, responsible for the relatively rigid wrapping of DNA in the NCP, and the formation of the stem in presence of the linker histone, as determined in the previous sections
- *internucleosome interactions*, such as steric hindrance between nucleosomes, which are likely to be important in dense fibers, but also more subtle effects like electrostatic interactions between nucleosomes, especially through the cationic histone tails, which proved to have an important role in chromatin folding [Shogren-Knaak et al. 2006, Arya and Schlick 2009].

It is not known, which of those interactions dominate the folding of the fibers. Is it the structure and elasticity of the stem, which discriminates between the possible helical arrangements of nucleosomes ? Or do the linkers simply adapt to whatever constraints follow from the packing and the interactions of the core particles [Depken and Schiessel 2009], thereby possibly frustrating the formation of a proper stem ?

The systematic modeling of the chromatin fiber thus requires estimates of these interactions, and in particular of the stem deformation and interaction free energies. Such estimates could be obtained from the exploration of the internal degrees of freedom of the stem through numerical simulations, and in particular the presumably complex electrostatic effects involved in the interaction of the flexible cationic C-terminus with the pair of DNA linkers (for a future work on this topic, see Section 3.5). Numerical simulation of fiber models is an active field of research, with contributions from a variety of groups using different types and levels of coarse-graining [Wedemann and Langowski 2002, Barbi" 2005, Kepper et al. 2008, Arya and Schlick 2009], including stem models which may share some features with ours [Wedemann and Langowski 2002, Stehr et al. 2008], and different computational methods [Barbi" 2005, Arya and Schlick 2007, Stehr et al. 2010]. We will not embark on the quantitative modeling of the fiber, which would be an entirely new project. Rather, we notice that a major obstacle in this field is the

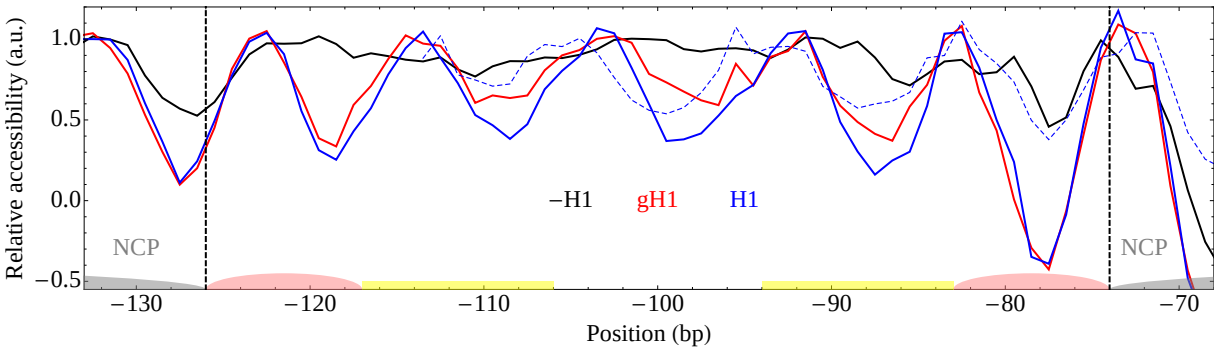


FIGURE 3.16 – Relative accessibility of the internal linker of the dinucleosome, obtained from the simultaneous treatment of the footprinting lanes in absence of H1 (black), in presence of gH1 (red) and of complete H1 (blue). The dotted blue line is the mononucleosome protection pattern, characteristic of a single undeformed stem. The vertical dashed lines indicate the limits of the linker DNA, and the pink and yellow semitransparent figures indicate the extent of gH1 and H1 tail respectively. The H1 maximally protected sites exhibits a shift of $\sim$ 1bp as compared to the mononucleosome case : this could be the consequence of (i) the superposition (or interference) of the protection patterns from the two stems ; (ii) a deformation of the stems, due to their mutual interaction ; (iii) an error in the positioning of the mononucleosome minima. These explanations are not mutually exclusive.

limited experimental data, which would be able to discriminate and refine the existing models. In the following, (i) we show that the analysis of polynucleosome footprinting gels may provide valuable information to test fiber models, and we begin this analysis with the available dinucleosome gels ; (ii) we discuss qualitatively the implications of our stem modeling for the chromatin fiber.

### 3.4.1 Analysis of dinucleosome gels

We consider the analysis of a new range of footprinting experiments, this time within nucleosome arrays. We reason that, just like the protection pattern on the mononucleosome linker gave bp resolved structural information on the stem structure, the protection pattern of the linker DNA buried in the array gives information on the (possibly deformed) stem and contacts between linker DNA of successive nucleosomes. Unfortunately, the experiments proved to be delicate for long arrays of nucleosomes, and we report here our analysis of dinucleosomes only. Trinucleosome gels were also analyzed, but the resolution did not allow a reliable identification of the bp on the linker. New experiments are being conducted on hexa- and dodecamers, and the present section must be understood as a first step (see Outlook section 3.5).

The analyzed gel (labeled gel 4 in Fig. 3.5) is the 53-bp long internal linker DNA of the dinucleosome. The resolved area is wider than for the mononucleosome, and the gel is less distorted, which allowed us to implement an improved version of the analysis procedure. We first used a home-made semi-automatic procedure to remove distortions present in the gel and thereby align the different lanes (see Section 3.4), so that the same pixel-bp transformation function can be used for all lanes simultaneously. As a result, using the same procedure as previously described, the resulting traces can be more accurately compared. In particular, we expect the identification procedure to be more reliable : “missing” a bp at some point would result in a simultaneous and sudden increase of all traces, which is not observed.

### (a) Stem interference

The result of the analysis is shown on Fig. 3.16. The raw data (after removing the deformations) can be found in Appendix, Fig. 3.20. Some features observed in the mononucleosome gels are again visible here, including the partial protection in absence of H1 (near the gH1 protection site), and in presence of gH1 ; the latter appears even stronger in this case.

In presence of H1, the characteristic $\sim$10-bp periodic protection pattern also presents a similar appearance as that of the mononucleosome (dashed line), albeit with some shift of the maximally protected sites. And yet the interpretation of the pattern is *very* different of the mononucleosome case. Indeed, the internal linker is involved in *two* stems simultaneously, one at either end. Thus, the observed protection pattern is in fact the *superposition* of two different, approximately periodic signals. If the two stems are equivalent, the *period* of these two signals can be expected to be the same, but there is no reason that these signals should be *in phase* : the relative phasing of the two signals is fixed by the linker length. Here, the linker length (53 bp) was chosen independently of this question, and it is therefore just by chance that the two protections are approximately in phase.

A schematic depiction of the corresponding structure is shown on Fig. 3.17A, right hand-side, under the hypothesis that the two stems are in their most favorable conformation, as determined in the mononucleosome. With the chosen linker length, the two stems do not overlap, and as a first approximation we may hypothesize that the root and trunk conformations are not affected by their interaction. The latter reduces to a simple volume exclusion in the crown, which is likely to conserve the phasing of the protection (as in the mononucleosome case) and thus the wavelength of both protections. In this case, the protection signal is quite analogous to an *interference pattern* between the traces of the two noninteracting stems. For specific linker lengths, one expects a vanishing of the oscillatory signal in the central region if the interference is destructive. On the other hand, the situation analogous to constructive interference is not possible : it would imply that the two external linkers lie on the same side of the protected internal linker, which is sterically impossible. For linker lengths where the two stems come close to one another, they experience an increasing mutual interaction, and the resulting deformations modify the protected sites on the internal linker, and thus the local wavelength : in this case the analogy to classical interference becomes inaccurate.

In our case, before any quantitative modeling, one can predict from the lack of a visible phase shift, that the two linkers must lie close to one another. To test if this statement is compatible with an undeformed stem, we will need to compute the corresponding protection pattern.

### (b) Precision of the analysis procedure

Before going into this structural and physical interpretation of the protection pattern, we would like to discuss the implications of the features described in the last paragraph, in terms of precision and reliability of the gel analysis.

In the analysis and subsequent modeling of the mononucleosomal stem, the whole procedure relied on the precise *absolute* positioning of the protected sites along the linker. This operation involved (i) the identification of the bp in the gels, and (ii) a delicate comparison with other gels where a combination of sequence- or molecular weight-dependent markers allowed to assign absolute positions to the bands. Although each of these steps has a sufficient resolving power, they are also prone to errors, both in the human treatment and in the semi-automatic counting procedures. Within our modeling scheme, an error of 1 bp in the identified site corresponds to

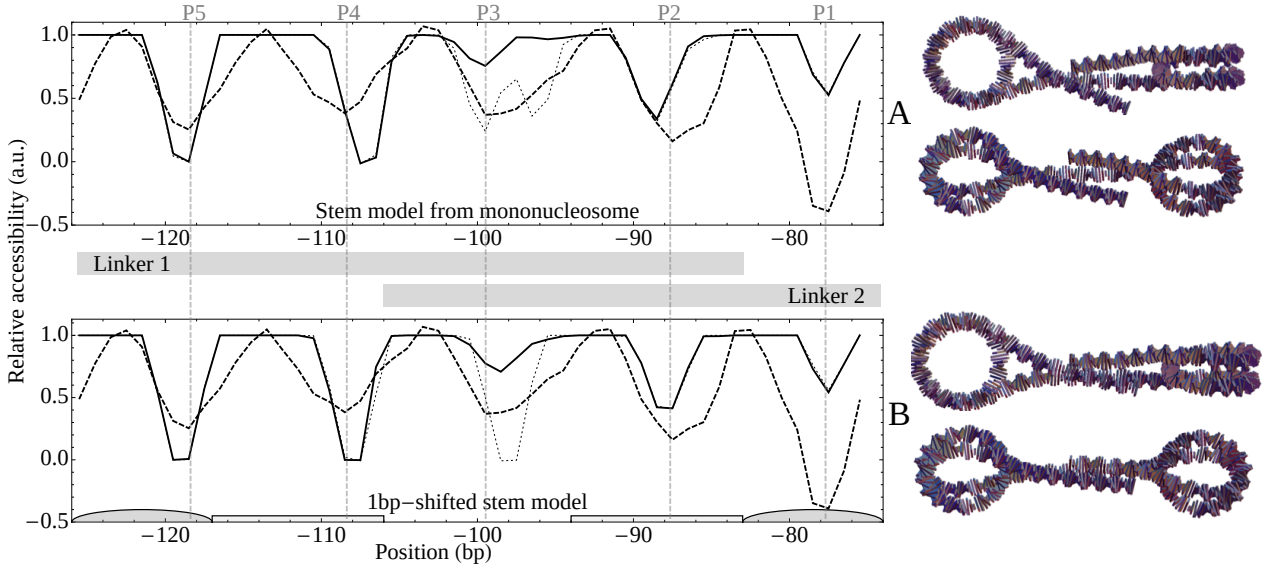


FIGURE 3.17 – Comparison of different stem models within a dinucleosome : **(A)** model obtained from the analysis of the mononucleosome, of elastic energy $2k_BT$ ; **(B)** distorted model, where the contacts are shifted of 1bp toward the NCP, of elastic energy $3.2k_BT$. **Left hand-side** Comparison of the model-derived ensemble-averaged relative accessibility (solid line) with the experimental signal (dashed line) and that of the corresponding rigid fully-protected structure (thin dotted line). The vertical gray dashed lines indicate the positions of the contacts (P1 to P3 can be related to their mononucleosome equivalents). The stem model obtained from the modeling of the mononucleosome predicts erroneously symmetrically shifted protected sites. A modified version of the stem corrects this effect, which may indicate a deformation of the trunk in the dinucleosome. The gray bars between the curves indicate the possible extensions of the external linkers. The gray ellipses and rectangles indicate the extensions of the gH1 and H1 tails respectively. **Right hand-side** Schematic depiction of the the corresponding shape of the dinucleosome structure, where the DNA is straight in the crown. Upper image : facing NCP1. Lower image : rotated 45° wrt the axis connecting the two NCP centers. The histones have not been depicted.

an over- or undertwisting of $\sim 34°$ of the linker DNA, with considerable consequences on the subsequent arrangement of nucleosomes in the fiber (see next section for a discussion).

Let us now consider the dinucleosome, and assume that the two stems are equivalent and symmetric : the latter hypothesis is certainly reasonable for linker lengths where the two stems are independent. Then, an error in the stem model (for instance, from a shift in the absolute positioning of the mononucleosome protection pattern) will result in a symmetric shift of the protected sites on either side of the linker : thus, a 1-bp error in the *absolute* positioning at the mononucleosome results in a 2-bp error in the *relative* distance between the two protected sites, which can be estimated much more reliably. Thus, if we use symmetric stem models to generate protection patterns, we can test these models by simply computing the distances between protected sites, and comparing it with the one observed in the gels, without any dependence on the absolute positioning. If we add the latter information, it then allows (i) to double check the positioning of the predicted protected sites and (ii) to register the positions of asymmetric protections, in the case where symmetric stem models are incompatible with the experimental signal.

### (c) Interpretation of the protection signal

The H1 protection pattern shown on Fig. 3.16 exhibits a shift of 1 bp, as compared to the mononucleosome pattern (dashed line). In the following, we discuss and test possible explanations.

**Interference between the protection patterns of the two noninteracting stems** Here "noninteracting" must be understood in the way described in the previous paragraph, *i.e.* no interaction between the roots and trunks. Fig. 3.17A shows that with the chosen linker length, the stems do not overlap and this hypothesis is plausible. The corresponding protection pattern is shown for the rigid fully protected structure (thin dotted line) and ensemble-averaged (solid). In the central region (P3), the interference pattern is clearly visible in the rigid structure ; however, one of the two peaks almost disappears because of thermal fluctuations. This asymmetric effect of fluctuations is compatible with the hypothesis of equivalent stems, because the internal DNA plays different roles in the two, once the entering linker, once the outgoing one : as a consequence, the protected sites of the analyzed strand are not at the same distance of the two NCPs.

On the other hand, the two sites protected by the tails (P2 and P4) exhibit no interference pattern, and thus the model predicts symmetrically erroneously shifted positions, corresponding (by construction of this stem model) to the minima of the mononucleosome signal. Note that this is an illustration of the effect discussed in the last paragraph, where a 1 bp shift in the absolute position results in two simultaneous and symmetrical shifts (or a 2-bp error in distance), giving more confidence in the processed signal. Here, we conclude that the signal is incompatible with the interference pattern of the two noninteracting stems.

We note however that we are reaching the limits of the predictive power of our coarse-grain protection : in the case of two interfering signals, the relative *amplitude* of the signals influences not only the amplitude of the global protection, but also the *position* of the minimum when the phase shift is small. Because our estimate of the degree of protection is only qualitative, we cannot estimate quantitatively the position of the minima due to such subtle effects. In this specific case however, the prediction is confirmed by the simple observation that only one of the two external linkers (indicated by the gray bars under the curve) is long enough to reach the considered site (P2/P4) : therefore, a symmetrical shift cannot be attributed to the effect of the opposed linker.

**Distorsion of the stems by mutual interaction** Since the protection pattern cannot be reproduced with the undeformed stem model from the mononucleosome, we suggest that the stem in the dinucleosome could be in an excited, deformed state. This deformation could result from the interaction between the two neighboring stems. We have generated new stem models in a procedure very similar to the relaxation of the mononucleosome, by imposing symmetrical contacts between the two linkers. However, in contrast to the latter case, we cannot simply place the springs at the minima of the dinucleosome signal, because they are the result of complex interactions between three linkers. Instead, we have generated a range of possible stems at different locations shifted by 0 or 1 bp from the mononucleosome minima (see Section 3.1.5), and computed the dinucleosome protection pattern for each of them. Not surprisingly, only the structures where the second contact, associated with the H1 tail, was shifted of 1 bp in the direction of the NCP were compatible with the minima P2 and P4, as shown in Fig. 3.17B, with

the corresponding structure of minimal energy : $3.2k_BT$ in the root and trunk regions.

The image of the corresponding structure on the right hand-side of the figure shows that this deformation brings the two external linkers very close to one another. This is rather surprising, since we would expect their repulsive forces to produce the opposite effect. As a putative explanation, the electrostatic effects might be more subtle, involving for instance rearrangements of the tails, which may impose much stronger constraints on the DNA than the relatively short-range repulsion between the linkers. Otherwise, an error in the positioning of the mononucleosome protections cannot be absolutely excluded, in which case this "deformed trunk" would in fact be the ground state. Another source of error is that the linker length is not well-defined in the experiments : in particular, if the nucleosomes incorporate only 145 bp as observed in the existing 601-NCP crystallographic structures Makde et al. [2010], Vasudevan et al. [2010], then the linker length would be 2 bp wider than estimated here, and therefore the mononucleosome protections would be shifted on either side, in a way qualitatively compatible with the observed (P2, P4). This hypothesis seems unlikely though, because in this case the gH1 protections (P1, P5) would be also shifted, which is not observed.

Finally, the deformed structure also incorrectly predicts the protection P3, in contrast to the original stem Fig. 3.17A. One of the constructed fully-protected structures has the correct "hybrid" combination of contacts, at an elastic cost of $\sim 30k_BT$ (mainly in the region between P2 and P3), however the wrong protection pattern Fig. 3.17B is recovered when the crown is allowed to fluctuate. The construction of a satisfactory model would thus require a more involved modeling ; on the other hand, as explained in the last paragraph, the level of detail of the coarse-grain protection estimation does not allow to quantitatively discriminate between the models in this central region where the "interference" effects may be important. Thus, we consider our deformed stem model as the most favorable in the limits of our modeling in the root and trunk regions (protections P1, P2, P4, P5). In the crown (P3), we estimate that the interactions between the linkers may be more important than for the mononucleosome, and the model of free fluctuations might be inaccurate.

### 3.4.2 Stem and fiber models

In absence of further gels of polynucleosome arrays, we cannot construct quantitatively improved fiber models. We will therefore stay at the qualitative level, and discuss how our results relate to the existing experimental and computational studies of the fiber.

#### (a) Extrapolation of the stem model to model fibers

Without any involved modeling of the interactions between nucleosomes, it is theoretically possible to extrapolate from the soft mononucleosome model obtained in the previous sections to build an ensemble of fiber conformations. Such a construction implicitly relies on the hypothesis that *the intra-nucleosome interactions dominate the fiber folding*, so that the trunk part of the stem remains undistorted. While most computational studies of the fiber explore the full conformational landscape of the fiber with limited resolution, our lower-scale coarse-grain models will allow us to discuss qualitatively the influence of the details in the stem modeling on the fiber conformations. We therefore build nucleosome arrays with (i) free linkers (H1-less nucleosomes) ; (ii) the mononucleosome stem model ; (iii) the deformed (overtwisted) stem compatible with the H1-tail protections of the dinucleosome gel.

Fig. 3.18A shows typical conformations of the *regular* fiber, *i.e.* in which the DNA is rigid in the crown, for a linker length (51 bp) where there are no steric clashes for the -H1 and the more compact mononucleosome stem fiber. For the dinucleosome stem however, the additional twist in the linker DNA modifies the relative orientation of the nucleosomes, resulting in severe steric clash. This is an illustration of the kind of dramatic consequence of a 1-bp shift in the gels, which we mentioned in the previous paragraph. On the other hand, we notice that the regular fiber with a 1-bp shorter linker (right hand-side image) has a much more similar shape. This is not an accident : the removal of a bp compensates for the overtwist of the linker in the stem, and the relative orientation of the successive nucleosome is roughly recovered. Next, we test if this qualitative effect remains within soft fiber.

For each linker length, we generate conformations of 20-nucleosome arrays through Monte Carlo sampling, in a way analogous to the construction of the trinucleosomes. However, in this case a simple Monte Carlo procedure is inefficient, because the probability of no overlap in an entire array is very small. We therefore implemented a naive chain growth algorithm [Rosenbluth and Rosenbluth 1955], which successfully generates fiber conformations by placing random nucleosome iteratively, and testing for overlaps at each step of the construction (see Section 3.1.7). The objective is to get a qualitative view of the type of conformations selected by our stem models, rather than a quantitatively accurate sampling of the fiber ensemble, which is anyway excluded by the prohibitive computing time in our implementation. Fig. 3.18B shows that while the -H1 fibers are more open and thus more easily constructed (dotted line), the two models of H1 arrays experience strong steric constraints that select 10-bp periodic linker lengths, with a slight phase shift consistent with the twist excess in the dinucleosome stem.

Thus, our minimalistic fiber models indicate that, at least qualitatively, inaccuracies in the detailed modeling of the stem, here induced by a shift in the protection pattern, may not strongly modify the energy landscape of the fiber, but rather shift it with respect to the linker length. This justifies the use of low-resolution, coarse-grain models of the stem in systematic models of the fiber, with the reverse consequence that there is a resolution limit in the "linker length" dimension. Note however that a similar uncertainty exists in fiber experiments, even when the position of the NCPs is well-controlled like in poly-601 arrays [Robinson et al. 2006] (see next section), because the number of bp in the NCP is not exactly known. In most crystals, they include only 146 or even 145 bp instead of 147. If the poly-601 nucleosomes in solution resemble the 145 bp crystals obtained with this sequence [Makde et al. 2010, Vasudevan et al. 2010], the linker length in these experiments may have a length $10n+2$ instead of a multiple of 10. As to the typical linker length *in vivo*, if a preferred 10 bp-periodicity has been recognized long ago [Lohr and Van Holde 1979, Widom 1992], the *phase* has been subject to controversy. Recent results in yeast, from both genome statistical analysis [Wang et al. 2008] and high-resolution nucleosome mapping [Brogaard and Widom 2012], seem to converge to favorite spacings $10n+5$ bp, which could be related to more open structures than the very compact "multiples of 10". But here again, there is an uncertainty of 2 bp.

We have discussed qualitatively the properties of the nucleosome arrays, as selected by our stem models. In the next paragraph, we compare these models with linker shapes inferred from the existing experimental data on the fiber.

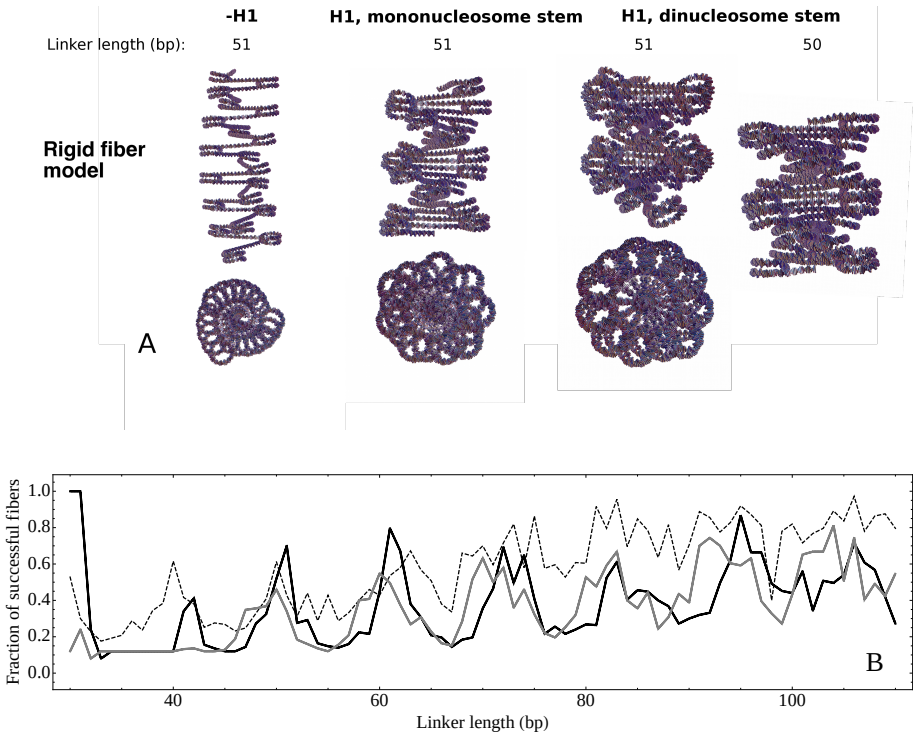


FIGURE 3.18 – Illustration of the nucleosome arrays selected by the mononucleosome stem and the overtwisted dinucleosome stem. (A) Images of regular arrays, for linker length 51bp. The -H1 array and the more compact mononucleosome stem array exhibit no steric clash, while the twist excess in the dinucleosome results in a very compact structure where the successive nucleosomes superpose. This is approximately resolved with a 1-bp shorter linker, which compensates the twist excess in the linker. (B) Fraction of successfully constructed fluctuating fibers, as estimated roughly from the chain growing algorithm (see text), for the -H1 array (thin dashed line), mononucleosome stem H1 (black), dinucleosome overtwisted stem H1 (gray). While the steric constraints are relatively loose in the -H1 array, they are more severe in presence of H1 and select 10-bp periodic linker lengths, with a phase shift consistent with the overtwist in the dinucleosome stem.

### (b) Qualitative comparison with linker DNA shapes inferred from model fibers

As mentioned, the experimental data on the chromatin fiber is rather limited. In [Robinson et al. 2006], Robinson *et al.* reported the properties of well-defined chromatin fibers reconstituted *in vitro* from poly-601 templates in presence of the linker histone H5. The templates were prepared for a range of well-controlled linker lengths (multiples of 10, from 30 to 90 bp). EM measurements showed that the fiber exhibits a transition between 60 and 70 bp : the diameter grows from $\sim 33 \pm 2.5$ nm to $\sim 44 \pm 3.5$ nm, and the nucleosome linear density from $\sim 11 \pm 1$ to $\sim 15 \pm 2$ nucleosomes/11nm. In [Wong et al. 2007], Wong *et al.* used DNA elastic models to infer the most favorable linker DNA paths from the properties of model fibers based on the same Robinson *et al* experiments. These structures, shown in Fig. 3.19 (A and C), belong to a remarkable variety of helix families [Wong et al. 2007]. Nevertheless, the authors argued that (in our language) the DNA conformation in the root is approximately conserved among the structures : *a posteriori*, this feature allows for a common asymmetric 3-contact binding mode of gH1/5, rather similar to the one inferred from our experiments. The only exception is the most favorable structure for the shortest linkers (30 bp or 15 bp for a *half-linker*), where the steric constraints seem indeed too strong to allow for the formation of the structure preferred

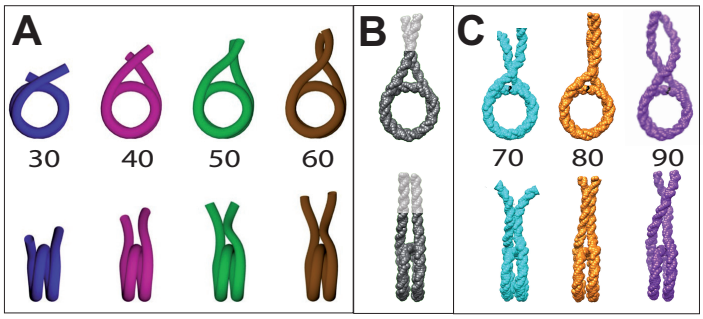


FIGURE 3.19 – Comparison of our mono-nucleosome stem structures to the inferred [Wong et al. 2007] linker conformations in model chromatin fibers reconstituted from poly-601 templates [Robinson et al. 2006]. View along the superhelical axis (top) and perpendicular (bottom). For the fiber conformations, only the first half of both linkers is shown. (A) Wong *et al.* most favorable structures for linker length 30, 40, 50 and 60 bp (from left to right), corresponding to a fiber diameter of 35 nm. Picture courtesy of Julien Mozziconacci. (B) Ground state of the mono-nucleosome stem : the root and trunk are those of the fully-protected structure, and the flexible crown or outer stem (shown in brighter colors) is straight. (C) Wong *et al.* most favorable structures for linker length 70, 80, 90 (from left to right), corresponding to a fiber diameter of 45 nm. The black dot indicates the dyad. Original pictures were not available : the linker conformations were rebuilt approximately, by hand, from the available data provided by Wong *et al.*. For all but the shortest linkers (30bp), the root part is approximately conserved among the structures, allowing for a common 3-contact asymmetrical binding mechanism of gH1 [Wong et al. 2007]. Longer linkers (50bp and beyond) adopt a conformation compatible with the formation of a trunk and additional stabilization by the H1 tail.

by gH1 and where Wong *et al.* propose an alternative binding mode. Otherwise, it is striking to see on Fig. 3.19A, how, with increasing linker length, the Wong *et al.* conformations reproduce larger and larger parts of the mono-nucleosomal stem shown in Fig. 3.19B, with very limited deformations. Linker lengths 40 bp and larger thus possibly allow for the formation of a (deformed) trunk and further stabilization by the H1 tail, whereas they exhibit substantial variations in the external part corresponding to our crown (Fig.3.19C).

At least qualitatively, the comparison of the two completely independent studies suggests a remarkable agreement between linker conformations required for achieving nucleosome packing in dense chromatin fibers and those stabilized by intra-stem interactions. Furthermore, the comparison illustrates the role of the polymorphic structure and hierarchic organization of the stem in stabilizing a wide variety of fibers.

## 3.5 Conclusion and outlook

### (a) Summary

While the structure of the nucleosome core particle is known with Ångström resolution, much less structural information is available for the soft parts of the nucleosome (linker DNA, linker histone H1/H5, histone tails), which strongly influence the structure of the chromatin fiber. The relevant structures are huge on the molecular scale and their intrinsic softness is a major obstacle to diffraction or NMR studies. On the other hand, they are too small to be imaged with the required resolution. Here we have presented a detailed report of our attempt to develop a three-dimensional, dynamical coarse-grain model of the nucleosomal stem formed by the

histone H1/H5 and the in- and outgoing linker DNA. We have combined available crystal and NMR structures (NCP and gH1 respectively) and the knowledge of the (sequence-dependent) B-DNA structure and elasticity with the results of our CEM and •OH footprinting experiments for carefully reconstituded model nucleosomes [Syed et al. 2010].

In the first step of our analysis we have shown that the expected protection pattern from the three-contact moDel [Fan and Roberts 2006] for the globular part of H1/H5 with the NCP is in excellent agreement with our analysis of the result of the footprinting experiments [Syed et al. 2010]. We then reasoned that the observed periodic protections on the linker DNA stem are a signature of DNA-DNA contacts, and that the precise positioning of the protected sites provides a valuable information. To model the stem structure, we have used a nanomechanical model of DNA, compatible with the qualitative character of the measured protection amplitude. We have aligned the linkers in space in a way that their mutual protection reproduces the measured accessibility profile and have assumed that the most likely structure has minimal DNA elastic energy. The resulting stem structure is shown in Figure 3.13C. The linkers come together $\sim$20 bases outside the core particle, slightly curving into a two-start superhelical stem with a large pitch of around 100-120 bp, and extending at least to bp 40 from the NCP. This structure has, as the core particle itself, a two-fold symmetry.

In the second step, we have developed a description of the stem in terms of an ensemble of fluctuating, partially-protected structures. We have shown that (i) transient contacts are sufficiently well defined to cause an experimentally observable partial protection if we assume a stem length of 20 or 24 bp, and (ii) the CEM pictures of reconstituted tri-nucleosomes are best reproduced by ensembles of fluctuating stems with a rigid part of 16 or 20 pbs. Combining these results, we therefore estimate that the rigid part of the stem incorporates $20 \pm 2$ bp of linker DNA. Interestingly, this corresponds to the extent of the linker DNA which is close enough to the globular part of H1 to directly interact with the (truncated) COOH-terminus of H1, which plays a crucial role in the stem formation. As a result of this analysis, the nucleosomal stem appears as a dynamic, polymorphic, hierarchically organized structure, composed of a "root" where gH1 binds to the first $\sim$ 10 bp of the DNA linkers, a "trunk" formed by the association of the subsequent $10 \pm 2$ bp with the cationic C-terminus of H1, and a flexible "crown" or outer stem where the branching linkers exhibit substantial fluctuations, while preserving well-defined preferential contacts.

Finally, we have discussed the consequences of our stem modeling wrt the folding of nucleosomes in the chromatin fiber, where the binding of H1 is associated to the selection of dense conformations. It is an open question, whether the dominant contribution of this selection comes from the *intra-nucleosomal* or the *inter-nucleosomal* interactions. In the latter case, the form of the stem *in vivo* may be relatively different from the structure found in our mononucleosome experiments : it is therefore not possible to simply extrapolate it to a polynucleosomal array. To get an insight into the matter, we analyzed footprinting gels from dinucleosomes, and we show that the signal is incompatible with the simple superposition of the protections as expected from the mononucleosome stem model : this result suggests that with the chosen linker length, the mutual interaction between the stems results indeed in a symmetric deformation, corresponding to an overtwist of 1 bp. We first showed that inaccuracies in the stem model, of the kind suggested by the dinucleosome analysis, do not prevent the exploration of fiber phase diagrams, but rather result in corresponding uncertainties in their coordinates, especially the linker length. Finally, we have compared our (thermal) ensemble of *mono*-nucleosome stem conformations to an ensemble of "most favorable" linker conformations [Wong et al. 2007] inferred

from experiments on reconstituted poly-601 *fibers* [Robinson et al. 2006]. Remarkably, for a wide range of linker lengths, the inferred linker conformations in the poly-601 fibers appear compatible with slightly deformed variants of the mono-nucleosomal stem. For shorter linkers, where the packing constraints become more severe, partial binding of H1 can still contribute to the stabilization of dense fibers. Finally, we hypothesize that the cooperative folding of dense chromatin fibers is facilitated by the hierarchical organisation of the stem and the tendency of local intra-stem interactions to stabilize linker conformations, which prevent non-local steric clashes between nucleosomes.

### (b) Outlook

**Analysis of polynucleosome gels** The analysis of the mononucleosome gels has provided detailed information on the stem, with an approximate bp-resolution. The same kind of analysis, carried on footprinting gels of linker DNA buried into polynucleosome arrays, may equally allow to infer structural information, both on the state of the stem and on the contacts between successive linkers. Such kind of quantitative information would be very useful to test and discriminate proposed fiber models, especially considering the limited existing amount of well-defined quantitative measurements on the fiber, even *in vitro*. The example of the dinucleosome gel showed that the interpretation of the signal is more delicate when the system becomes more complex, with several interacting linkers : therefore, it may be necessary to refine our coarse-grained protection function, as well as integrate data obtained with different linker lengths. Experiments on 12-nucleosome arrays are in preparation.

**Modeling the stem electrostatics** In our modeling, we included the electrostatic interactions only implicitly, by the means of artificial springs enforcing contacts at the experimentally observed positions. Then, in the construction of the soft structure, they were treated according to the following hypothesis :

- in the region where the cationic tail is in direct contact with the DNA (trunk), the free energy (i) has its ground state in the conformation observed in the maximally-protected structure ; (ii) has an infinite stiffness wrt the deformations of the DNA : the system is therefore frozen
- in the region beyond (crown), where there can be no attraction, the two linker branches do not experience any electrostatic interaction, except for an excluded volume

Clearly, both assumptions are extremely coarse. In the frame of our study, they are justified by the qualitative character of the protection signal, and by the absence of external forces. In the fiber however, we noticed that the inter-nucleosome interactions could generate strong constraints on the stem, in which case not only the crown, but also the trunk could be deformed. It is therefore important to estimate the stem deformation free energy landscape.

This is a difficult problem, because of the important amount of positive (tail, cations) and negative (phosphate groups of the two neighboring linkers) charges in the confined space between the two DNA arms in the trunk : at such concentrations, the simplified models valid in the crown [Schiessel 2002] may break down. This is made even worse by the specific helicoidal arrangement of the negative charges and the important steric effects in their vicinity, which are known to result in subtle effects, such as “zipper-motives” [Kornyshev" 1999].

As a solution, one could consider all-atomic MD simulations. However, the simulation time may be considerable for properly sampling the tail rearrangements and the cationic movements

in a very rugged energy landscape.

Instead, we propose to consider intermediate models such as the Poisson-Boltzmann derived approaches, and in particular a recently proposed model where the ion shapes are taken into account [Koehl et al. 2009], and thus the steric effects are approximately included. For instance, the estimated density of charges and water molecules around a DNA oligomer reproduces the experimentally observed helicoidal arrangements. The limited computational cost may then allow to compute the free energy for a large number of DNA conformations, under the hypothesis that the ion equilibration is faster than the DNA large-scale movements, and thus compute a coarse-grained free energy function wrt the DNA coordinates. A validation of the model on simpler DNA systems is underway.

# Appendix

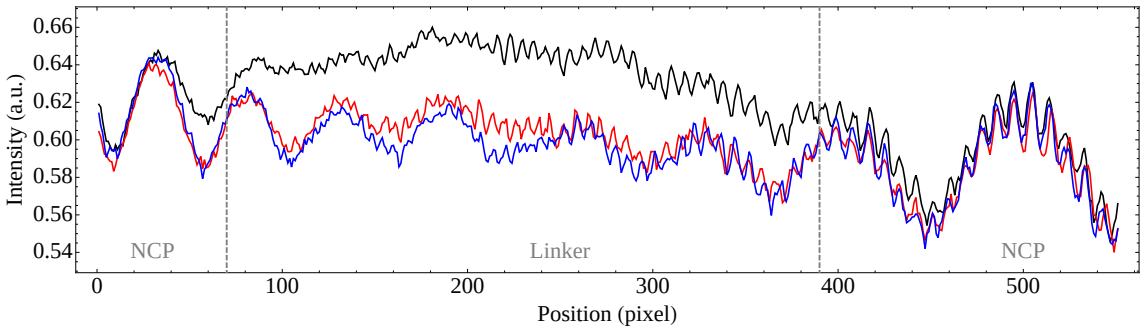


FIGURE 3.20 – Intensity of the -H1 (black), gH1 (red) and H1 (blue) lanes in the dinucleosome linker gel (gel 4, see Fig. 3.5), after correction for gel distorsions. The superposition of the different signals on the NCP (right part of the curves) confirm that the procedure aligned the corresponding bp of the different lanes with bp resolution.

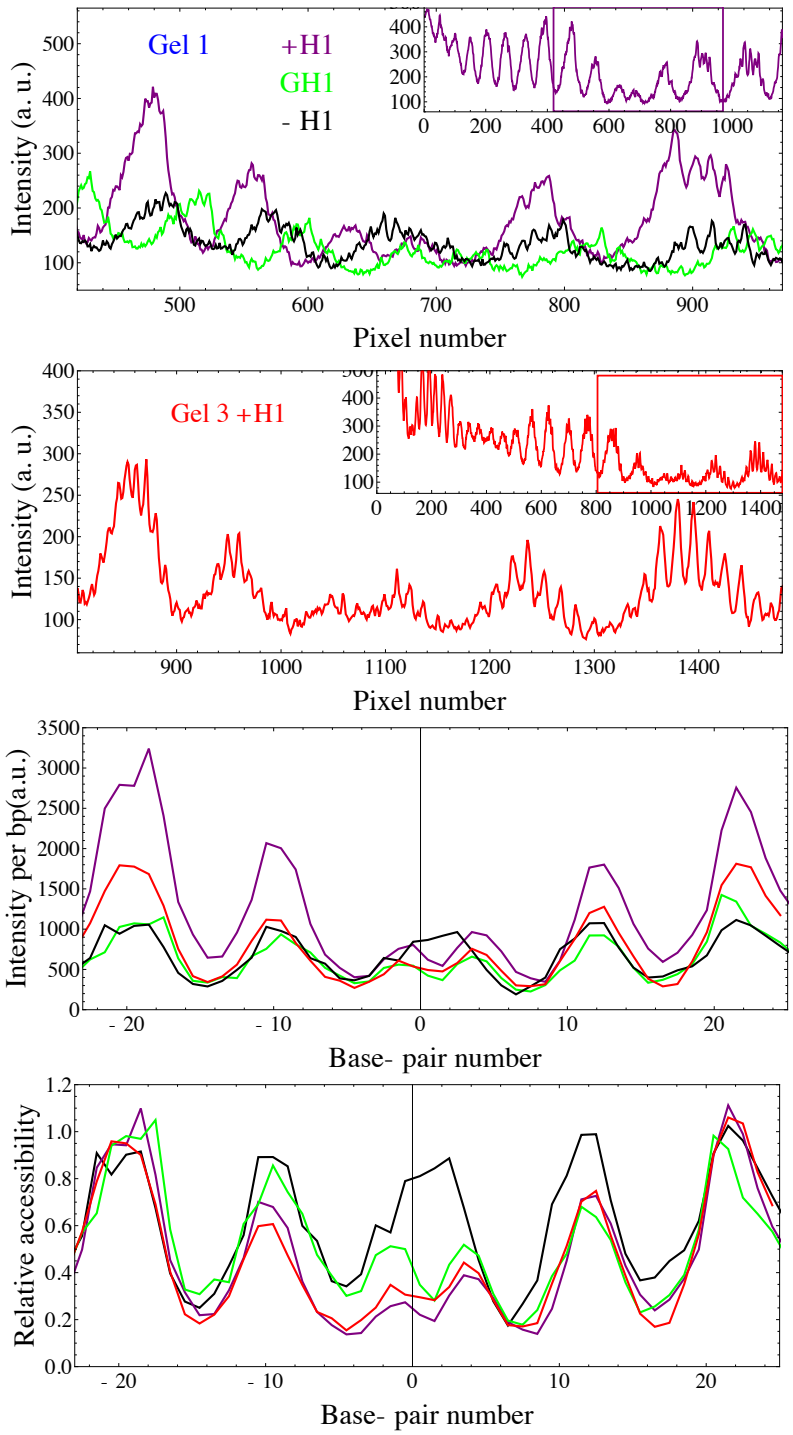


FIGURE 3.21 – Post-processing of the foot-printing gels in the dyad region of the NCP. **Upper panels** : Available raw data for mononucleosome (gel 1) and dinucleosome (gel 3, same as the black trace of Fig. 2, left panel) respectively (see Fig. 3.5). In each case, the main figure shows the resolved region of the gel. Insets show the whole signal. **Third panel :** Intensity per bp. **Bottom panel :** relative accessibilities.

# Bibliographie


Abascal, J. and Vega, C. (2005). A general purpose model for the condensed phases of water : Tip4p/2005. *The Journal of chemical physics*, 123 :234505. 23

Andersen, H. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics*, 72 :2384. 33

Angelov, D., Lenouvel, F., Hans, F., Muller, C. W., Bouvet, P., Bednar, J., Moudrianakis, E. N., Cadet, J., and Dimitrov, S. (2004). The histone octamer is invisible when nf-kappab binds to the nucleosome. *J Biol Chem*, 279(41) :42374–42382. 116

Angelov, D., Vitolo, J., Mutskov, V., Dimitrov, S., and Hayes, J. (2001). Preferential interaction of the core histone tail domains with linker dna. *Proceedings of the National Academy of Sciences*, 98(12) :6599. 18

Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M., and Scipioni, A. (2000). A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. *Biophysical journal*, 79(2) :601–613. 88

Arents, G. and Moudrianakis, E. N. (1995). The histone fold : a ubiquitous architectural motif utilized in dna compaction and protein dimerization. *Proc Natl Acad Sci U S A*, 92(24) :11170–11174. 110

Arya, G. and Schlick, T. (2006). Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model. *Proc Natl Acad Sci U S A*, 103(44) :16236–16241. 110, 133

Arya, G. and Schlick, T. (2007). Efficient global biopolymer sampling with end-transfer configurational bias monte carlo. *J Chem Phys*, 126(4) :044107. 134

Arya, G. and Schlick, T. (2009). A tale of tails : how histone tails mediate chromatin compaction in different salt and linker histone environments. *J Phys Chem A*, 113(16) :4045–59. 18, 110, 134

Balasubramanian, B., Pogozelski, W. K., and Tullius, T. D. (1998). Dna strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the dna backbone. *Proc Natl Acad Sci U S A*, 95(17) :9738–9743. 112, 119, 126, 127

Bao, Y., White, C., and Luger, K. (2006). Nucleosome core particles containing a poly (da·dt) sequence element exhibit a locally distorted dna structure. *Journal of molecular biology*, 361(4) :617–624. 108

Barbi", M. (2005). How the chromatin fiber deals with topological constraints. *Physical Review E*, 71(3). 134

Barrat, J. and Hansen, J. (2003). *Basic concepts for simple and complex liquids*. Cambridge

Univ Pr. 74

Battistini, F., Hunter, C., Moore, I., and Widom, J. (2012). Structure-based identification of new high affinity nucleosome binding sequences. *Journal of Molecular Biology*. 88

Becker, N. (2007). *Sequence dependent elasticity of DNA*. PhD thesis, Max-Planck-Institut fur Physik komplexer Systeme. 25, 28

Becker, N. B. and Everaers, R. (2007). From rigid base pairs to semiflexible polymers : coarse-graining dna. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76(2 Pt 1) :021923. 25, 26, 29, 30, 67, 68, 101, 118

Becker, N. B. and Everaers, R. (2009a). Dna nanomechanics : how proteins deform the double helix. *J Chem Phys*, 130(13) :135102. 28, 89

Becker, N. B. and Everaers, R. (2009b). Dna nanomechanics in the nucleosome. *Structure*, 17(4) :579–589. 28, 88, 95, 96, 99, 100, 109, 129

Becker, N. B., Wolff, L., and Everaers, R. (2006). Indirect readout : detection of optimized subsequences and calculation of relative binding affinities using different dna elastic potentials. *Nucleic Acids Res*, 34(19) :5638–5649. 25, 118, 129

Bednar, J., Horowitz, R. A., Grigoryev, S. A., Carruthers, L. M., Hansen, J. C., Koster, A. J., and Woodcock, C. L. (1998). Nucleosomes, linker dna, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc Natl Acad Sci U S A*, 95(24) :14173–14178. 12, 20, 111

Berendsen, H., Postma, J., Van Gunsteren, W., DiNola, A., and Haak, J. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81 :3684. 32, 44

Beveridge, D. L., Barreiro, G., Byun, K. S., Case, D. A., Cheatham, 3rd, T. E., Dixit, S. B., Giudice, E., Lankas, F., Lavery, R., Maddocks, J. H., Osman, R., Seibert, E., Sklenar, H., Stoll, G., Thayer, K. M., Varnai, P., and Young, M. A. (2004). Molecular dynamics simulations of the 136 unique tetranucleotide sequences of dna oligonucleotides. i. research design and results on d(cpg) steps. *Biophys J*, 87(6) :3799–813. 23, 25, 42, 44

Bharath, M. M. S., Chandra, N. R., and Rao, M. R. S. (2003). Molecular modeling of the chromatosome particle. *Nucleic Acids Res*, 31(14) :4264–74. 133

Biswas, M., Wocjan, T., Langowski, J., and Smith, J. (2012). Dna bending potentials for loop-mediated nucleosome repositioning. *EPL (Europhysics Letters)*, 97 :38004. 87

Blosser, T., Yang, J., Stone, M., Narlikar, G., and Zhuang, X. (2009). Dynamics of nucleosome remodelling by individual acf complexes. *Nature*, 462(7276) :1022–1027. 19

Brochier-Armanet, C., Forterre, P., et al. (2007). Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea*, 2(2) :83–93. 36

Brogaard, K. K. and Widom, J. (2012). A map of nucleosome positions in yeast at base-pair resolution ; nature. *Nature*. 36, 140

Brown, D. T., Izard, T., and Misteli, T. (2006). Mapping the interaction surface of linker histone h1(0) with the nucleosome of native chromatin in vivo. *Nat Struct Mol Biol*, 13(3) :250–255. 110, 118, 119, 126, 127, 133

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity resca-

ling. *The Journal of chemical physics*, 126 :014101. 33

Calladine, C. R. and Drew, H. R. (1997). *Understanding DNA : the molecule and how it works*. Academic Press, San Diego, 2nd edition. 9, 13, 118

Case, D., Cheatham III, T., Darden, T., Gohlke, H., Luo, R., Merz Jr, K., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. (2005). The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16) :1668–1688. 43

Cerf, C., Lippens, G., Ramakrishnan, V., Muyldermans, S., Segers, A., Wyns, L., Wodak, S. J., and Hallenga, K. (1994). Homo- and heteronuclear two-dimensional nmr studies of the globular domain of histone h1 : full assignment, tertiary structure, and comparison with the globular domain of histone h5. *Biochemistry*, 33(37) :11079–11086. 118

Chakravarthy, S. and Luger, K. (2006). The histone variant macro-h2a preferentially forms hybrid nucleosomes. *Journal of Biological Chemistry*, 281(35) :25522–25531. 108

Chandler, D. (1987). Introduction to modern statistical mechanics. *Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10 : 0195042778. ISBN-13 : 9780195042771*, 1. 27

Cheatham, T. E., Cieplak, P., and Kollman, P. A. (1999). A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. *J Biomol Struct Dyn*, 16(4) :845–862. 43, 93

Chevereau, G., Palmeira, L., Thermes, C., Arneodo, A., and Vaillant, C. (2009). Thermodynamics of intragenic nucleosome ordering. *Phys Rev Lett*, 103(18) :188103. 36

Chua, E., Vasudevan, D., Davey, G., Wu, B., and Davey, C. (2012). The mechanics behind dna sequence-dependent properties of the nucleosome. *Nucleic Acids Research*. 89

Clapier, C., Chakravarthy, S., Petosa, C., Fernández-Tornero, C., Luger, K., and Müller, C. (2008). Structure of the drosophila nucleosome core particle highlights evolutionary constraints on the h2a-h2b histone dimer. *Proteins : Structure, Function, and Bioinformatics*, 71(1) :1–7. 108

Clark, D. J. and Thomas, J. O. (1986). Salt-dependent co-operative interaction of histone h1 with linear dna. *J Mol Biol*, 187(4) :569–580. 112

Claudet, C., Angelov, D., Bouvet, P., Dimitrov, S., and Bednar, J. (2005). Histone octamer instability under single molecule experiment conditions. *J Biol Chem*, 280(20) :19958–19965. 110

Collins, F., Lander, E., Rogers, J., Waterston, R., and Conso, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945. 9

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19) :5179–5197. 22, 93

Cuesta-Lopez, S., Angelov, D., and Peyrard, M. (2009). Adding a new dimension to dna melting curves. *EPL (Europhysics Letters)*, 87 :48009. 36

Cuesta-López, S., Menoni, H., Angelov, D., and Peyrard, M. (2011). Guanine radical chemistry reveals the effect of thermal fluctuations in gene promoter regions. *Nucleic Acids Research*, 39(12) :5276–5283. 69

Cui, F. and Zhurkin, V. B. (2009). Distinctive sequence patterns in metazoan and yeast nucleosomes : implications for linker histone binding to at-rich and methylated dna. *Nucleic Acids Res*, 37(9) :2818–29. 133

Dang, L. (1995). Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether : a molecular dynamics study. *Journal of the American Chemical Society*, 117(26) :6954–6960. 43

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald : An n log (n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98 :10089. 32, 44

Dauxois, Peyrard, and Bishop (1993). Entropy-driven dna denaturation. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 47(1) :R44–R47. 36

Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W., and Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol*, 319(5) :1097–1113. 18, 99, 108, 116, 118, 133

Davey, G., Wu, B., Dong, Y., Surana, U., and Davey, C. (2010). Dna stretching in the nucleosome facilitates alkylation by an intercalating antitumour agent. *Nucleic acids research*, 38(6) :2081–2088. 108

de la Barre, A. E., Angelov, D., Molla, A., and Dimitrov, S. (2001). The n-terminus of histone h2b, but not that of histone h3 or its phosphorylation, is essential for chromosome condensation. *EMBO J*, 20(22) :6383–6393. 110

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J., and Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium aquifex aeolicus. *Nature*, 392(6674) :353–358. 36

Deniz, Ö., Flores, O., Battistini, F., Pérez, A., Soler-López, M., and Orozco, M. (2011). Physical properties of naked dna influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC genomics*, 12(1) :489. 69, 88

Depken, M. and Schiessel, H. (2009). Nucleosome shape dictates chromatin fiber structure. *Biophys J*, 96(3) :777–84. 134

Dickerson, R. (1989). Definitions and nomenclature of nucleic acid structure components. *Nucleic acids research*, 17(5) :1797–1803. 16, 44

Dixit, S. B., Beveridge, D. L., Case, D. A., Cheatham, 3rd, T. E., Giudice, E., Lankas, F., Lavery, R., Maddocks, J. H., Osman, R., Sklenar, H., Thayer, K. M., and Varnai, P. (2005). Molecular dynamics simulations of the 136 unique tetranucleotide sequences of dna oligonucleotides. ii : sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys J*, 89(6) :3721–40. 25, 42, 44

Dixon, N. and Kornberg, A. (1984). Protein hu in the enzymatic replication of the chromosomal origin of escherichia coli. *Proceedings of the National Academy of Sciences*, 81(2) :424. 19

Dorigo, B., Schalch, T., Bystricky, K., and Richmond, T. J. (2003). Chromatin fiber folding : requirement for the histone h4 n-terminal tail. *J Mol Biol*, 327(1) :85–96. 17, 110

Dorigo, B., Schalch, T., Kulangara, A., Duda, S., Schroeder, R. R., and Richmond, T. J. (2004). Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science*, 306(5701) :1571–1573. 11, 20

Eberhart, J. and II, V. P. (1985). The thermodynamic limit of superheat of water. *Journal of Colloid and Interface Science*, 107(2) :574 – 575. 40, 62

Fan, L. and Roberts, V. A. (2006). Complex of linker histone h5 with the nucleosome and its implications for chromatin packing. *Proc Natl Acad Sci U S A*, 103(22) :8384–8389. 110, 118, 119, 126, 127, 133, 143

Fang, H., Clark, D. J., and Hayes, J. J. (2012). Dna and nucleosomes direct distinct folding of a linker histone h1 c-terminal domain. *Nucleic Acids Res*, 40(4) :1475–1484. 20, 111

Fermi, E., Pasta, J., and Ulam, S. (1955). Studies of nonlinear problems. Technical report, I, Los Alamos Scientific Laboratory Report No. LA-1940. 31

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J., and van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, 143(2) :212–224. 21, 111

Flaus, A. and Owen-Hughes, T. (2001). Mechanisms for atp-dependent chromatin remodelling. *Curr Opin Genet Dev*, 11(2) :148–154. 10, 19, 20

Flyvbjerg, H. and Petersen, H. (1989). Error estimates on averages of correlated data. *Journal of Chemical Physics*, 91(1) :–. 45, 49, 73

Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation : From Algorithms to Applications*. Academic Press, San Digeo, CA. 31, 33, 34

Gansen, A., Valeri, A., Hauger, F., Felekyan, S., Kalinin, S., Tóth, K., Langowski, J., and Seidel, C. (2009). Nucleosome disassembly intermediates characterized by single-molecule fret. *Proceedings of the National Academy of Sciences*, 106(36) :15308–15313. 18

Geggier, S., Kotlyar, A., and Vologodskii, A. (2011). Temperature dependence of dna persistence length. *Nucleic Acids Res*, 39(4) :1419–26. 11, 37, 68

Gelbart, W., Bruinsma, R., Pincus, P., and Parsegian, V. (2000). Dna-inspired electrostatics. *Physics today*, 53 :38. 14

Gonzalez, O. and Maddocks, J. (2001). Extracting parameters for base-pair level models of dna from molecular dynamics simulations. *Theoretical Chemistry Accounts : Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 106(1) :76–82. 45

Heine, M. and Chandra, S. B. C. (2009). The linkage between reverse gyrase and hyperthermophiles : a review of their invariable association. *J Microbiol*, 47(3) :229–34. 36

Hoover, W. et al. (1985). Canonical dynamics : equilibrium phase-space distributions. *Physical Review A*, 31(3) :1695–1697. 33

Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations : Tip4p-ew. *J Chem Phys*, 120(20) :9665–78. 23, 43, 45

Horowitz-Scherer, R. A. and Woodcock, C. L. (2006). Organization of interphase chromatin. *Chromosoma*, 115(1) :1–14. 111

Hurst, L. and Merchant, A. (2001). High guanine–cytosine content is not an adaptation to high temperature : a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 268(1466) :493–497. 36

Ioshikhes, I. P., Albert, I., Zanton, S. J., and Pugh, B. F. (2006). Nucleosome positions predicted through comparative genomics. *Nat Genet*, 38(10) :1210–1215. 19

Jones, E., Oliphant, T., and Peterson, P. (2001). Scipy : Open source scientific tools for python. 44

Jost, D. and Everaers, R. (2008). A unified poland-scheraga model of oligo- and polynucleotide dna melting : salt effects and predictive power. *Biophys J*. 36

Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000). Archaeal adaptation to higher temperatures revealed by genomic sequence of thermoplasma volcanium. *Proc Natl Acad Sci U S A*, 97(26) :14257–14262. 36

Kepper, N., Foethke, D., Stehr, R., Wedemann, G., and Rippe, K. (2008). Nucleosome geometry and internucleosomal interactions control the chromatin fiber conformation. *Biophys J*, 95(8) :3692–705. 134

Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in drosophila melanogaster. *Nature*, 471(7339) :480–485. 21, 111

Kimball, A., Guo, Q., Lu, M., Cunningham, R. P., Kallenbach, N. R., Seeman, N. C., and Tullius, T. D. (1990). Construction and analysis of parallel and antiparallel holliday junctions. *J Biol Chem*, 265(25) :15348. 127

Koehl, P., Orland, H., and Delarue, M. (2009). Solvation of ion pairs : The poisson-langevin model. In *2009 International Conference on Signal Processing Systems*, pages 917–923. IEEE. 23, 145

Koopmans, W., Brehm, A., Logie, C., Schmidt, T., and van Noort, J. (2007). Single-pair fret microscopy reveals mononucleosome dynamics. *Journal of Fluorescence*, 17(6) :785–795. 133

Koopmans, W., Buning, R., Schmidt, T., and Van Noort, J. (2009). spfret using alternating excitation and fcs reveals progressive dna unwrapping in nucleosomes. *Biophysical journal*, 97(1) :195–204. 18

Kornberg, R. D. (1974). Chromatin structure : a repeating unit of histones and dna. *Science*, 184(139) :868–871. 10

Kornyshev", A. (1999). Electrostatic zipper motif for dna aggregation. *Physical Review Letters*, 82(20) :4138–4141. 111, 144

Kruithof, M., Chien, F.-T., Routh, A., Logie, C., Rhodes, D., and van Noort, J. (2009). Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. *Nat Struct Mol Biol*, 16(5) :534–540. 133

Kulic, I. M. and Schiessel, H. (2003a). Chromatin dynamics : nucleosomes go mobile through twist defects. *Phys Rev Lett*, 91(14) :148103. 18, 19, 88, 133

Kulic, I. M. and Schiessel, H. (2003b). Nucleosome repositioning via loop formation. *Biophys J*, 84(5) :3197–3211. 18, 19, 36, 87, 133

Kulic, I. M. and Schiessel, H. (2004). Dna spools under tension. *Phys Rev Lett*, 92(22) :228101. 133

Lankas, F., Sponer, J., Langowski, J., and Cheatham, T. E. r. (2003). Dna basepair step deformability inferred from molecular dynamics simulations. *Biophys J*, 85(5) :2872–2883. 25, 43, 118, 129

Lavery, R., Moakher, M., Maddocks, J., Petkeviciute, D., and Zakrzewska, K. (2009). Conformational analysis of nucleic acids revisited : Curves+. *Nucleic acids research*, 37(17) :5917–5929. 26, 44

Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, T. r., Dixit, S., Jayaram, B., Lankas, F., Laughton, C., Maddocks, J. H., Michon, A., Osman, R., Orozco, M., Perez, A., Singh, T., Spackova, N., and Sponer, J. (2010). A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in b-dna. *Nucleic Acids Res*, 38(1) :299–313. 24, 25, 38, 42, 46, 52

Leach, A. R. (2001). *Molecular modelling : principles and applications*. Addison-Wesley Longman Ltd. 22, 32

Li, B., Carey, M., and Workman, J. (2007). The role of chromatin during transcription. *Cell*, 128(4) :707–719. 19

Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005). Rapid spontaneous accessibility of nucleosomal dna. *Nat Struct Mol Biol*, 12(1) :46–53. 18, 87, 123, 133

Lohr, D. and Van Holde, K. E. (1979). Organization of spacer dna in chromatin. *Proc Natl Acad Sci U S A*, 76(12) :6326–30. 140

Lowary, P. T. and Widom, J. (1998). New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol*, 276(1) :19–42. 36, 88, 93, 112

Lu, X.-J. and Olson, W. K. (2003). 3dna : a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, 31(17) :5108–5121. 26, 118, 120

Luger, K. and Hansen, J. C. (2005). Nucleosome and chromatin fiber dynamics. *Curr Opin Struct Biol*, 15(2) :188–196. 111

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648) :251–260. 89, 92, 108

Makarov, V. L., Dimitrov, S. I., Tsaneva, I. R., and Pashev, I. G. (1984). The role of histone h1 and non-structured domains of core histones in maintaining the orientation of nucleosomes within the chromatin fiber. *Biochem Biophys Res Commun*, 122(3) :1021–1027. 110

Makde, R. D., England, J. R., Yennawar, H. P., and Tan, S. (2010). Structure of rcc1 chromatin factor bound to the nucleosome core particle. *Nature*, 467(7315) :562–6. 17, 89, 93, 108, 133, 139, 140

Mangenot, S., Leforestier, A., Vachette, P., Durand, D., and Livolant, F. (2002). Salt-induced conformation and interaction changes of nucleosome core particles. *Biophys J*, 82(1 Pt

1) :345–356. 110, 133

Marguet, E. and Forterre, P. (1994). Dna stability at temperatures typical for hyperthermophiles. *Nucleic acids research*, 22(9) :1681–1686. 36

Matsumoto, A. and Olson, W. (2002). Sequence-dependent motions of dna : a normal mode analysis at the base-pair level. *Biophysical journal*, 83(1) :22–41. 25

Mergell, B., Everaers, R., and Schiessel, H. (2003). Nucleosome interactions in chromatin : Fiber stiffening and hairpin formation. *Phys. Rev. E*, 70 :011915. 21

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21 :1087. 34

Meyer, S., Becker, N. B., Syed, S. H., Goutte-Gattat, D., Shukla, M. S., Hayes, J. J., Angelov, D., Bednar, J., Dimitrov, S., and Everaers, R. (2011). From crystal and nmr structures, footprints and cryo-electron-micrographs to large and soft structures : nanoscale modeling of the nucleosomal stem. *Nucleic Acids Res*, 39(21) :9139–54. 113, 117

Michor, F., Liphardt, J., Ferrari, M., and Widom, J. (2011). What does physics have to do with cancer ? *Nature Reviews Cancer*, 11(9) :657–670. 10

Milani, P., Chevereau, G., Vaillant, C., Audit, B., Haftek-Terreau, Z., Marilley, M., Bouvet, P., Argoul, F., and Arneodo, A. (2009). Nucleosome positioning by genomic excluding-energy barriers. *Proc Natl Acad Sci U S A*, 106(52) :22257–62. 36

Montel, F., Fontaine, E., St-Jean, P., Castelnovo, M., and Faivre-Moskalenko, C. (2007). Atomic force microscopy imaging of swi/snf action : Mapping the nucleosome remodeling and sliding. *Biophysical Journal*, 93(2) :566 – 578. 111, 123, 133

Moreira, D. and López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, 7(4) :306–311. 35

Morozov, A. V., Fortney, K., Gaykalova, D. A., Studitsky, V. M., Widom, J., and Siggia, E. D. (2009). Using dna mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Res*, 37(14) :4707–22. 88

Muhlbacher, F., Schiessel, H., and Holm, C. (2006). Tail-induced attraction between nucleosome core particles. *Phys Rev E Stat Nonlin Soft Matter Phys*, 74(3 Pt 1) :031919. 110, 133

Muthurajan, U. M., Bao, Y., Forsberg, L. J., Edayathumangalam, R. S., Dyer, P. N., White, C. L., and Luger, K. (2004). Crystal structures of histone sin mutant nucleosomes reveal altered protein-dna interactions. *EMBO J*, 23(2) :260–271. 108

Nelson, D., Lehninger, A., and Cox, M. (2008). *Lehninger principles of biochemistry*. WH Freeman. 19, 36

Nosé, S. (1984). A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2) :255–268. 33

Olson, W., Bansal, M., Burley, S., Dickerson, R., Gerstein, M., Harvey, S., Heinemann, U., Lu, X., Neidle, S., Shakked, Z., et al. (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of molecular biology*, 313(1) :229–237. 16, 44

Olson, W. and Zhurkin, V. (2011). Working the kinks out of nucleosomal dna. *Current Opinion in Structural Biology*. 19, 90, 93, 95, 97

Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., and Zhurkin, V. B. (1998). Dna sequence-dependent deformability deduced from protein-dna crystal complexes. *Proc Natl Acad Sci U S A*, 95(19) :11163–11168. 24, 89, 118, 129

Ong, M., Richmond, T., and Davey, C. (2007). Dna stretching and extreme kinking in the nucleosome core. *Journal of molecular biology*, 368(4) :1067–1074. 108

Ouldridge, T., Louis, A., and Doye, J. (2010). Structural, mechanical and thermodynamic properties of a coarse-grained dna model. *Arxiv preprint arXiv :1009.4480*. 21

Pachov, G. V., Gabdoulline, R. R., and Wade, R. C. (2011). On the structure and dynamics of the complex of the nucleosome and the linker histone. *Nucleic Acids Res*, 39(12) :5255–63. 133

Pastor, N., Weinstein, H., Jamison, E., and Brenowitz, M. (2000). A detailed interpretation of oh radical footprints in a tbp-dna complex reveals the role of dynamics in the mechanism of sequence-specific binding. *J Mol Biol*, 304(1) :55–68. 119, 120

Pearlman, D., Case, D., Caldwell, J., Ross, W., Cheatham, T., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1) :1–41. 43

Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T., Laughton, C., and Orozco, M. (2007). Refinement of the amber force field for nucleic acids : improving the description of [alpha]/[gamma] conformers. *Biophysical journal*, 92(11) :3817–3829. 43, 93

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13) :1605–1612. 14, 118, 119, 126

Polach, K., Widom, J., et al. (1995). Mechanism of protein access to specific dna sequences in chromatin : a dynamic equilibrium model for gene regulation. *Journal of molecular biology*, 254(2) :130–149. 18, 19

Polach, K., Widom, J., et al. (1996). A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *Journal of molecular biology*, 258(5) :800–812. 18

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007). *Numerical Recipes 3rd Edition : The Art of Scientific Computing*. Cambridge University Press. 51, 52, 101

Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L., and Sweet, R. M. (1993). Crystal structure of globular domain of histone h5 and its implications for nucleosome binding. *Nature*, 362(6417) :219–223. 111, 118

Richmond, T. J. and Davey, C. A. (2003). The structure of dna in the nucleosome core. *Nature*, 423(6936) :145–150. 17, 19, 88, 95, 97, 101, 104

Robinson, P. J. J., An, W., Routh, A., Martino, F., Chapman, L., Roeder, R. G., and Rhodes, D. (2008). 30 nm chromatin fibre decompaction requires both h4-k16 acetylation and linker histone eviction. *J Mol Biol*, 381(4) :816–825. 110

Robinson, P. J. J., Fairall, L., Huynh, V. A. T., and Rhodes, D. (2006). Em measurements define the dimensions of the "30-nm" chromatin fiber : evidence for a compact, interdigitated

structure. *Proc Natl Acad Sci U S A*, 103(17) :6506–6511. 11, 21, 140, 141, 142, 144

Rochman, M., Postnikov, Y., Correll, S., Malicet, C., Wincovitch, S., Karpova, T. S., McNally, J. G., Wu, X., Bubunenko, N. A., Grigoryev, S., and Bustin, M. (2009). The interaction of nsbp1/hmgn5 with nucleosomes in euchromatin counteracts linker histone-mediated chromatin compaction and modulates transcription. *Mol Cell*, 35(5) :642–56. 111

Rosenbluth, M. and Rosenbluth, A. (1955). Monte carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics*, 23 :356. 140

Ryckaert, J., Ciccotti, G., and Berendsen, H. (1977). title = "numerical integration of the cartesian equations of motion of a system with constraints : molecular dynamics of n-alkanes",. *Journal of Computational Physics*, 23(3) :327–341. 32, 44

Sahu, G., Wang, D., Chen, C., Zhurkin, V., Harrington, R., Appella, E., Hager, G., and Nagaich, A. (2010). P53 binding to nucleosomal dna depends on the rotational positioning of dna response element. *Journal of Biological Chemistry*, 285(2) :1321. 19

Sanner, M. F., Olson, A. J., and Spehner, J. C. (1996). Reduced surface : an efficient way to compute molecular surfaces. *Biopolymers*, 38(3) :305–320. 119, 121

SantaLucia, Jr, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A*, 95(4) :1460–5. 36, 62

Saunders, N. F. W., Thomas, T., Curmi, P. M. G., Mattick, J. S., Kuczek, E., Slade, R., Davis, J., Franzmann, P. D., Boone, D., Rusterholtz, K., Feldman, R., Gates, C., Bench, S., Sowers, K., Kadner, K., Aerts, A., Dehal, P., Detter, C., Glavina, T., Lucas, S., Richardson, P., Larimer, F., Hauser, L., Land, M., and Cavicchioli, R. (2003). Mechanisms of thermal adaptation revealed from the genomes of the antarctic archaea methanogenium frigidum and methanococcoides burtonii. *Genome Res*, 13(7) :1580–1588. 36

Schalch, T., Duda, S., Sargent, D. F., and Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436(7047) :138–141. 11, 20, 134

Schiessel, H. (2002). How short-ranged electrostatics controls the chromatin structure on much larger scales. *Europhys. Lett.*, 58(1) :140. 144

Schiessel, H. (2003). The physics of chromatin. *J. Phys. Cond. Mat.*, 15 :699–774. 9, 18, 20, 36, 110, 111, 134

Schiessel, H., Widom, J., Bruinsma, R. F., and Gelbart, W. M. (2001). Polymer reptation and nucleosome repositioning. *Phys. Rev. Lett.*, 86(19) :4414–4417. 133

Schumacker, R. and Lomax, R. (2004). *A beginner's guide to structural equation modeling*, volume 1. Lawrence Erlbaum. 54

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104) :772–778. 19, 36

Shintomi, K., Iwabuchi, M., Saeki, H., Ura, K., Kishimoto, T., and Ohsumi, K. (2005). Nucleosome assembly protein-1 is a linker histone chaperone in xenopus eggs. *Proc Natl Acad Sci U S A*, 102(23) :8210–8215. 112

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R., and Peterson, C. L. (2006). Histone h4-k16 acetylation controls chromatin structure and protein interactions. *Science*,

311(5762) :844–847. 17, 110, 134

Shukla, M. S., Syed, S. H., Montel, F., Faivre-Moskalenko, C., Bednar, J., Travers, A., Angelov, D., and Dimitrov, S. (2010). Remosomes : Rsc generated non-mobilized particles with approximately 180 bp dna loosely associated with the histone octamer. *Proc Natl Acad Sci U S A*, 107(5) :1936–41. 19, 133

Sinden, R. (1994). *DNA structure and function*. Academic Pr. 15

Stehr, R., Kepper, N., Rippe, K., and Wedemann, G. (2008). The effect of internucleosomal interaction on folding of the chromatin fiber. *Biophysical Journal*, 95(8) :3677 – 3691. 134

Stehr, R., Schöpflin, R., Ettig, R., Kepper, N., Rippe, K., and Wedemann, G. (2010). Exploring the conformational space of chromatin fibers and their stability by numerical dynamic phase diagrams. *Biophys J*, 98(6) :1028–37. 134

Strahs, D., Barash, D., Qian, X., and Schlick, T. (2003). Sequence-dependent solution structure and motions of 13 tata/tbp (tata-box binding protein) complexes. *Biopolymers*, 69(2) :216–243. 119, 120

Suto, R., Clarkson, M., Tremethick, D., and Luger, K. (2000). Crystal structure of a nucleosome core particle containing the variant histone h2a. z. *Nature Structural & Molecular Biology*, 7(12) :1121–1124. 108

Syed, S. H., Goutte-Gattat, D., Becker, N., Meyer, S., Shukla, M. S., Hayes, J. J., Everaers, R., Angelov, D., Bednar, J., and Dimitrov, S. (2010). Single-base resolution mapping of h1-nucleosome interactions and 3d organization of the nucleosome. *Proc Natl Acad Sci U S A*, 107(21) :9620–9625. 110, 112, 125, 126, 127, 128, 129, 133, 143

Tachiwana, H., Kagawa, W., Osakabe, A., Kawaguchi, K., Shiga, T., Hayashi-Takanaka, Y., Kimura, H., and Kurumizaka, H. (2010). Structural basis of instability of the nucleosome containing a testis-specific histone variant, human h3t. *Proceedings of the National Academy of Sciences*, 107(23) :10454. 108

Tansey, M. R. and Brock, T. D. (1972). The upper temperature limit for eukaryotic organisms. *Proc Natl Acad Sci U S A*, 69(9) :2426–2428. 35

Theodorakopoulos, N. and Peyrard, M. (2012). Base pair openings and temperature dependence of dna flexibility. *Phys Rev Lett*, 108(7) :078104. 11, 37, 66, 67, 68

Tolstorukov, M., Colasanti, A., McCandlish, D., Olson, W., and Zhurkin, V. (2007). A novel roll-and-slide mechanism of dna folding in chromatin : implications for nucleosome positioning. *Journal of molecular biology*, 371(3) :725–738. 90, 95

Tomschik, M., Zheng, H., Van Holde, K., Zlatanova, J., and Leuba, S. (2005). Fast, long-range, reversible conformational fluctuations in nucleosomes revealed by single-pair fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9) :3278. 18

Travers, A., Muskhelishvili, G., Thompson, J., Travers, A., Muskhelishvili, G., and Thompson, J. (2012). Dna information : from digital code to analogue structure. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 370(1969) :2960–2986. 38, 69

Tremethick, D. J. (2007). Higher-order structures of chromatin : the elusive 30 nm fiber. *Cell*, 128(4) :651–654. 20, 111

Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. (1999). The packing density in proteins : standard radii and volumes. *J Mol Biol*, 290(1) :253–266. 120

Tsunaka, Y., Kajimura, N., Tate, S., and Morikawa, K. (2005). Alteration of the nucleosomal dna path in the crystal structure of a human nucleosome core particle. *Nucleic acids research*, 33(10) :3424–3434. 108

van Holde, K. and Zlatanova, J. (2007). Chromatin fiber structure : Where is the problem now ? *Semin Cell Dev Biol*, 18(5) :651–658. 11

Vasudevan, D., Chua, E. Y. D., and Davey, C. A. (2010). Crystal structures of nucleosome core particles containing the '601' strong positioning sequence. *J Mol Biol*, 403(1) :1–10. 89, 93, 108, 139, 140

Vega, C. and Abascal, J. (2005). Relation between the melting temperature and the temperature of maximum density for the most common models of water. *The Journal of chemical physics*, 123 :144504. 23

Voltz, K., Trylska, J., Calimet, N., Smith, J., and Langowski, J. (2012). Unwrapping of nucleosomal dna ends : A multiscale molecular dynamics study. *Biophysical Journal*, 102(4) :849–858. 88

Wang, J.-P., Fondufe-Mittendorf, Y., Xi, L., Tsai, G.-F., Segal, E., and Widom, J. (2008). Preferentially quantized linker dna lengths in saccharomyces cerevisiae. *PLoS Comput Biol*, 4(9) :e1000175. 36, 140

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356) :737–738. 9, 13, 14, 24

Wedemann, G. and Langowski, J. (2002). Computer simulation of the 30-nanometer chromatin fiber. *Biophys. J.*, 82 :2847–2859. 88, 110, 134

White, C., Suto, R., and Luger, K. (2001). Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *The EMBO journal*, 20(18) :5207–5218. 108

White, M. and Bell, S. (2002). Holding it together : chromatin in the archaea. *TRENDS in Genetics*, 18(12) :621–626. 19

Widom, J. (1992). A relationship between the helical twist of dna and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc Natl Acad Sci U S A*, 89(3) :1095–9. 140

Wolfram Research, I. (2008). *Mathematica Edition : Version 7.0*. Wolfram Research, Inc., Champaign, Illinois. 25, 101, 116

Wong, H., Victor, J.-M., and Mozziconacci, J. (2007). An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PLoS ONE*, 2(9) :e877. 133, 141, 142, 143

Woodcock, C. L. and Dimitrov, S. (2001). Higher-order structure of chromatin and chromosomes. *Curr Opin Genet Dev*, 11(2) :130–135. 10, 20

Woodcock, C. L., Grigoryev, S. A., Horowitz, R. A., and Whitaker, N. (1993). A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. *Proc Natl Acad Sci U S A*, 90(19) :9021–9025. 110

Wu, B., Mohideen, K., Vasudevan, D., and Davey, C. (2010). Structural insight into the sequence dependence of nucleosome positioning. *Structure*, 18(4) :528–536. 108

Wu, C., Bassett, A., and Travers, A. (2007). A variable topology for the 30-nm chromatin fibre. *EMBO Rep*, 8(12) :1129–1134. 10, 11

Wu, C. and Travers, A. (2005). Relative affinities of dna sequences for the histone octamer depend strongly upon both the temperature and octamer concentration. *Biochemistry*, 44(43) :14329–14334. 38

Zhang, Y., Moqtaderi, Z., Rattner, B., Euskirchen, G., Snyder, M., Kadonaga, J., Liu, X., and Struhl, K. (2009). Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol*. 36

Zhaxybayeva, O., Swithers, K., Lapierre, P., Fournier, G., Bickhart, D., DeBoy, R., Nelson, K., Nesbø, C., Doolittle, W., Gogarten, J., et al. (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the thermotogales. *Proceedings of the National Academy of Sciences*, 106(14) :5865–5870. 35

Zhou, Y. B., Gerchman, S. E., Ramakrishnan, V., Travers, A., and Muyldermans, S. (1998). Position and orientation of the globular domain of linker histone h5 on the nucleosome. *Nature*, 395(6700) :402–405. 110, 118, 119, 126, 127, 133, 134

Zlatanova, J., Leuba, S. H., and van Holde, K. (1998). Chromatin fiber structure : morphology, molecular determinants, structural transitions. *Biophys. J.*, 74 :2554–2566. 110

# Glossary

| | |
|---|---|
| AA | Amino Acid |
| AFM | Atomic Force Microscopy |
| bp | base-pair |
| bps | base-pair step |
| $\underline{C}$ | covariance matrix |
| CEM | Cryo-Electron Micrograph |
| dh | double-helix |
| dof | degree(s) of freedom |
| dsDNA | double-stranded DNA |
| FRET | Fluorescence Resonance Energy Transfer |
| fs | femtosecond ($10^{-15}$ s) |
| gH1 | the globular domain of the linker histone H1 |
| H1 | linker histone H1 |
| $\underline{K}$ | stiffness matrix |
| lhs | left hand-side |
| $l_p$ | persistence length |
| MC | Monte Carlo |
| MD | Molecular Dynamics |
| NCP | Nucleosome Core Particle |
| nm | nanometer |
| NMR | Nuclear Magnetic Resonance |
| ns | nanosecond ($10^{-9}$ s) |
| ps | picosecond ($10^{-12}$ s) |
| q | 6-dimensional coordinate vector |
| rbp | rigid base-pair |
| rhs | right hand-side |
| SHL | superhelical location |
| ss | single-strand |
| wlc | worm-like chain |